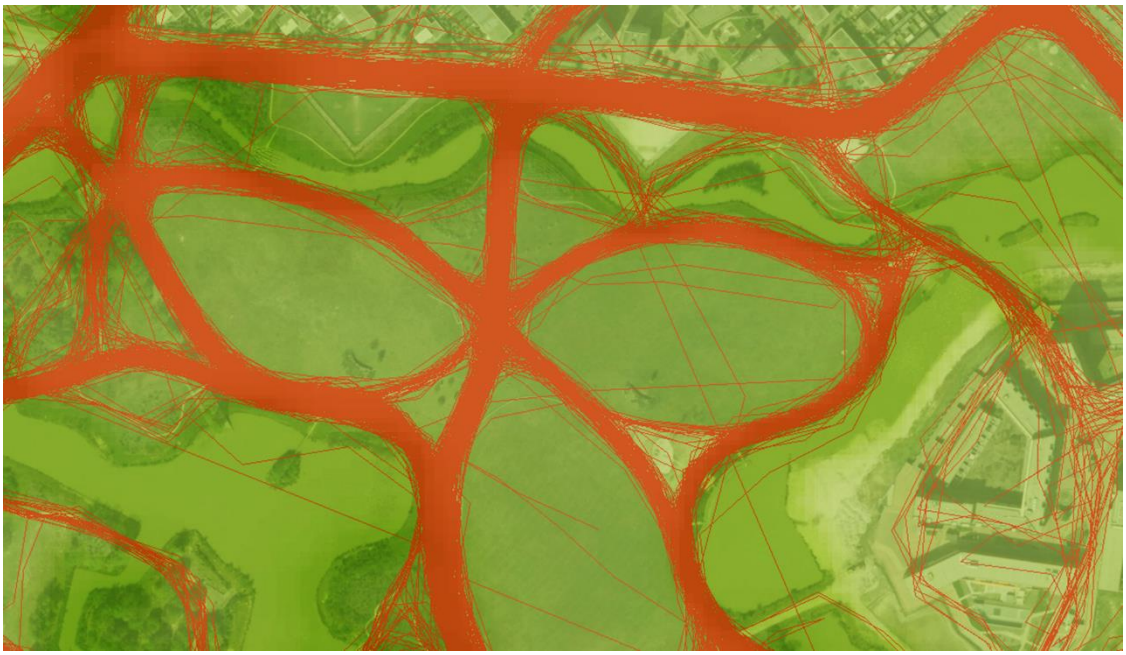
GIMA
Geographical Information Management and Applications

Running the Tracks

Contextual Influence Modeling



Simon Groen
3857611 / s6026044

Utrecht University
Geographical Information Management
and Applications

Supervisor: Ir. Edward Verbree
Professor: Prof.dr.ir. P.J.M. van Oosterom

Front page figure: Runner tracks and natural areas spatial influence in the city park of Meerhoven, Eindhoven (Google, 2019)

Colophon

GIMA MSc Thesis
16-08-2019

This master thesis is commissioned by the University of Utrecht, Wageningen University & Research, University of Twente, and the Delft University of Technology as part of the Geographical Information Management and Applications (GIMA) Master of Science.

GIMA Professor:
Prof.dr.ir. P.J.M. van Oosterom

Supervisor:
Ir. Edward Verbree

Contact Information:
Simon Groen
Nassaustraat 4A
3583 XE Utrecht
groensimon93@gmail.com

UU student number: 3857611
ITC student number: s6026044

Acknowledgements

I want to start with expressing my deepest gratitude towards my supervisor at the start of this master thesis, Simon Scheider. He had helped me a lot at the start of the thesis with setting up the subject matter, providing relevant data and knowledge, and teaching me the basics of programming with Python. Sadly, conducting the thesis research proved to be my Achilles heel. The size and own commitment is not something that is easy to deal with for me personally when the work is not done in a project group. This has led to numerous delays in finishing this thesis. At one point, Simon Scheider could understandably no longer guide me through this process, after which Ir. Edward Verbree stepped in as a second supervisor. I am very thankful he gave me and this subject a second chance of wrapping everything up.

Throughout my life I have invested a lot of time into different sports, be it tennis, climbing, football, basketball, running or fitness. Choosing a topic for my master thesis related to physical activity, therefore, seems a logical direction for my interests to go within the confounds of the GIMA subject matter. I hope that my passion for physical activity in general shines through in this thesis report you are about to read and consequently I hope you will have a good time reading it.

Abstract

Research into physical activity done in the built environment has become increasingly more important the last few decades, as an ever growing part of people in western countries participate in it. Among them running as activity has become important, being one of the most performed sports in the Netherlands. This increase in popularity sparked a similar increase in attention given by scientific research to the topic. These studies often do not yet take geographical information into account. This master thesis aims to participate in filling up this gap of implementation by researching if the spatial influences on a runner can be modeled using geographical information.

To do this, nine influence factors are gathered from scientific literature; running surface, verbal harassment, street lighting, motorized vehicles, cyclists, natural areas, sound pollution, air pollution and variety in surroundings. For these factors, influence modeling methods are composed to map the spatial influences based on a runner being on an influence source, being in close proximity or receiving influences from multiple sources around the runner. By enriching 200 GPS tracks with this influence information for each GPS measurement. To try and validate the methods and influence factors, the results are statistically tested against the amount of runner activity per neighborhood in a research area around the Dutch city of Eindhoven. A multiple regression analysis is performed with the nine influence factors per GPS measurements as independent variables.

The performance of the regression model, however, seems poor, as the relation between the influence factors and the amount of runner activity in a neighborhoods share a moderate, but significant, relation. Causes for this were found in both the uncertainties in the modeling methods, as this is an explorative study, as limitations in the data that could be used to model the factors. Further research into combinations of influence factors, research subject sizes and modeling methods is needed to assess if the groundwork this thesis achieves in researching the spatial influences on runners is to be utilized further.

Table of Contents

List of Abbreviations

σ	-	Standard deviation
CBS	-	Statistics Netherlands
CV	-	Coefficient of Variation
dB	-	Decibel
df	-	Degrees of Freedom
ERP	-	Extended Research Proposal
GI	-	Geographical Information
GIS	-	Geographical Information System
GIScience	-	Geographical Information Science
GPS	-	Global Positioning System
HMM	-	Hidden Markov Model
MAUP	-	Modifiable Areal Unit Problem
MCA	-	Multi Criteria Analysis
OSM	-	OpenStreetMap
P	-	Influence Probability
PABE	-	Physical Activity and Built Environment
RIVM	-	Netherlands National Institute for Public Health and the Environment
SEI	-	Stimulating Environment Index
UGCoP	-	Uncertain Geographical Context Problem
V	-	Influence Value

Chapter 1
Introduction

1.1 Research Context

"Running is associated with desirable lifestyle change"

Kluitenberg, et al., 2013, p. 1

Running as a physical activity for everyone has seen an immense growth in popularity over the past 60 years. Halfway through the last century, it was still mostly a way of exercise done on athletics tracks specifically designed for it. From that time on, running has experienced two waves which contributed greatly to the rise in popularity and acceptance of leisure-time running, as this was seen as a waste of time before. The first wave started in the 1960ies when running on the streets, in parks or in nature became more accepted. This cumulated in the increased offer of marathons in the 70s and the integration of leisure-time runners into the marathons in the 80s. Thereafter, the popularity of running stagnated until the start of the 21st century. Running and sport in general became an integral part of life as women also began participating on a large scale in it. This is seen as the second wave that increased the popularity of running as physical activity, a wave we are still in (Scheerder & Breedveld, 2015).

During the first wave of increasing popularity, the Dutch government started to actively promote running as a way to a healthy lifestyle by for example building 2-4 kilometer runs in forests and parks (Reiling & Dolders, 2015). Things like governmental promotion have resulted in an increased amount of Dutch citizens that participate in physical activity in accordance to the Dutch guidelines. In the past fifteen years alone, this number has risen from little over half to three-quarter of the adult population of the Netherlands (Bernaards, et al., 2015). On the flipside of this positive trend is that sedentary time (being still, except for sleeping) per day per person is also growing amongst the Dutch population during both work and free time. Reasons for this growth are the changing nature of jobs and the increased roles of screen technologies (e.g. laptops, tablets and smartphones) in our lives, making the trend hard to counter (Bernaards, et al., 2011). The increasing amount of adults exercising in accordance to the Dutch guidelines concerning physical activity can therefore be seen as a reaction to this sedentary lifestyle.

Not only physical activity in general in the Netherlands has risen since the start of the 21st century, but running as well. In 2012, 13% of the Dutch population between 6 and 80 years old participated in

running. Only hiking (14%), swimming (18%) and fitness/aerobics (22%) are practiced by a larger share of the population (Hover, 2013). 13% means that 1,9 million people in the Netherlands participated in running in 2012 (Veiligheid.nl, 2014). This percentage does not include jogging and trimming, however, two forms of physical activity that are closely related to running, but are measured separately nowadays. The most recent study into a joined percentage is that 22% of the Dutch population between 6 and 80 years old participated in either of the three physical activities in 2007 (Statistics Netherlands [CBS], 2011). One conclusion to be drawn from these facts is that stimulating the population to take up running can have a great effect on the sedentary lifestyle. This is in line with the opening quote of this paragraph from Kluitenberg et al. (2013). Of the three forms of physical activity more popular than running, only hiking is also done outside of confined and specific sport areas (e.g. swimming pool, gym). This means that both hiking and running are more difficult to effectively stimulate, as they are not confined to specific 'arenas'. All the public space could be used to perform these activities. Stimulation therefore comes from the environment runners run through, a broad and complex context. Formulating a methodology that could help with identifying the optimal environment for runners based on its influence on them is the central theme of this thesis.

1.2 Research Field and Relevance

Before talking about the contribution this research can give to science, it is important to define in which scientific field it is positioned. Reiling and Dolders (2015) wrote their master thesis on *"Designing a runner friendly city"*, a subject quite similar to this master thesis in the sense that they too look at what makes an environment runner friendly. However, this thesis does not focus on actual design but rather on the factors that make up a runner friendly environment and how to define their influence on a runner. Reiling and Dolders (2015) defined the scientific field to which they were contributing as Physical Activity and Built Environment, or PABE for short. However, an aspect of human geography mostly left out in their thesis is the incorporation of geographical data in the research methodology. Their approach was mostly qualitative, with expert interviews being the major contribution to their findings. The incorporation of geographical data instead of a qualitative approach is the aspect that differentiates this thesis from theirs (see Chapter 3). This does not mean that the scientific field to which it contributes is different, however. It means that while this research has subject that can be seen as a PABE study, it shares a greater link with a different field of science, Geographical Information Science [GIScience]. This field is situated between geography, computer science and information science. The studying geographic information by means of computational approaches and data structures with the right tools, Geographical Information Systems [GIS] (Goodchild, 2010).

Harris et al. (2013) looked at 318 PABE studies and their approaches and scopes. Of those, 191 were discovery focused, 79 were reviews of previous studies, only 38 focused on formulating theory and methods for studying PABE, while 6 were focused on delivery (having results that are directly usable by policy makers). As becomes clear when reading this thesis, it has a primary focus on formulating a computational methodology (see paragraph 1.3). This means only a small portion of PABE research up to 2013 shares the general approach of this master thesis. In addition, of the 38 studies focusing on formulating theory and methods, not a lot provide a link between PABE and GIScience. There is, however, a second way in which this research separates itself from other PABE research, which is its focus on runners. Ettema (2015) points out that most studies concerning physical activity and its relation

to the built environment are about hikers or cyclists. There is only a small pool of research done about runners and even less include geographical information in it. Most articles in the small pool about runners use a qualitative approach.

This research addresses a gap in the scientific research that exists in the PABE field between studies focusing on why runners run certain tracks or why they avoid them (see Ettema, 2015) and studies that discuss and analyze public space using geographical data (see Forsyth, 2000; Ostermann & Timpf, 2009; Cebrecos, e.a., 2016). As said before, the optimal hiking environment has been researched more extensively. As Ettema and Smajic (2015) point out, however, comparing the needs of hikers and runners is not that straightforward. While some elements in the public space stimulate or hinder both hikers and runners, this cannot be said about all of them and even if the effect is the same, the strength or importance often still differs. Additionally, a computational methodology for hikers is lacking. Therefore, the need for contemporary research that makes use of different kinds of geographic data about runners is important. These difficulties of comparing hikers and runners are explored in detail in paragraph 2.1.

The purpose of this research, however, is not merely scientific. The results of it can also be important to parties involved in planning public space, both rural and urban. With the aforementioned increases in inactivity among most age groups (Bernaards, et al., 2011), it is important that public space stimulates physical activity. The fact that running is one of the most common forms of physical activity outside of designated sport fields (see paragraph 1.1) makes the topic of this thesis socially relevant. A better understanding of the influence of the built environment on runners and a methodological approach is needed to address the inactivity problem. Enhancing public space to stimulate activeness is already part of the policy concerning sport of certain municipalities. The Municipality of Eindhoven, for example, aims to remove physical hindrances such as fences from parks. Furthermore, they are searching for the right aesthetic, lighting and surface for sport activities in these areas (Municipality of Eindhoven, 2015).

1.3 Problem Statement

As said in paragraph 1.2, a gap in the PABE literature is noticeable when it comes to the use of geographical data in research about how spatial factors influence runners. As mentioned before, this master thesis has a primary focus on formulating a computational methodology. This methodology can be seen as reaching certain research objectives. Together, these should encompass the thought process behind getting the right results in this thesis.

1. *To determine which factors in the public space might have an impact on the activeness of runners, be it negative or positive.*

The first research objective encompasses the theoretical input necessary for the other parts the research. PABE literature with previous findings is used to determine which factors are important to integrate into the methodology. The expected result is a definitive set of attractivity factors and obstacles in the running environment.

2. *To formulate methods of spatial influence modeling and assign them to the influence factors.*

The spatial factors need to be made operational. This is done by using different geographical datasets and determining to what extent they represent these factors. Each factor can be modelled using different datasets or attributes within the same dataset. The ways in which an influence factor can be operationalized is discussed in Chapter 3.

3. *To assess the effect of the modifiable areal unit problem and uncertain geographical context problem on the spatial influences.*

When working with multiple geographical datasets at the same time, two problems arise. The first and also most well documented is the modifiable areal unit problem. Choosing inadequate areal units can cause unforeseen bias in the results. Furthermore, a problem that is often overlooked when working with geographical data is that of the uncertain geographic context. This problem is especially apparent when working with the effects of area-based attributes on individuals behaviors (Kwan, 2012), as is done in this research with the effects of factors on runners. For this reason, assessing an adequate geographical context for the influence factors is the third research objective. Both problems are discussed extensively in paragraph 2.4.

4. *To formulate a methodology for enriching GPS tracks with the spatial influences using the available geographical data.*

To utilize the operationalized spatial influences, a method needs to be formulated on how to assess how much these influence the runners. This is done by enriching GPS tracks of runners with these influences. The result of this objective is GPS tracks with information on the extent to which they are influenced by each factor, but the focus is on how this is to be done. Methods and techniques for this are explored in greater detail in Chapter 3.

5. *To determine quality of the research methodology through validation.*

The final research objective entails the validation of the research results. As mentioned with the fourth research objective, the results are a quantitative measurement on how the influence factors influence runners. To assess the fit for use of the methodology, these results have to be validated in some way. This is done by determining to what extent each factor contributes to the presence of runners.

These five research objectives encompass the scope of the research process. Reaching these objectives results in the completion of this process. To get the why of the whole process, the five objectives are combined into three research objectives, which will be explained in the next paragraph.

1.4 Research Questions

To reach these objectives, three research questions have been formulated. The first addresses geographical theory, while the second and third have a methodological focus.

RQ1: Which spatial factors in the running environment are known to have a positive influence on the activeness of runners and which factors are known to have a negative influence on it?

This first research question entails the theoretical exploration of relevant literature from the PABE scientific field. By using a broad selection of scientific research, it is aimed to achieve a definitive list of environmental influence factors which have either a positive or negative influence on the runner.

RQ2: How can the spatial influence factors be operationalized with the available geographical data?

When the spatial influence factors are known, they need to be modelled and with them their context. This means determining how approximately each influence influences a runner spatially. This process brings with it its own problems and uncertainties, which are explained in more detail in paragraph 2.4.

RQ3: To what extent can the operationalized spatial influences and research methodology be validated?

With the operationalized influence factors and the enrichment of the routes, the initial results of the research methodology are gathered. To really say something about the usefulness of the methods used, it is important to try and validate them.

1.5 Case Study Introduction

As evident from the research objectives and questions, the spatial influences on runners during a run and how they can be modeled is the empirical focus of this master thesis. To facilitate this research focus, the right case study needs to be selected. The Technical University of Eindhoven and Fontys University of Applied Sciences have provided the researcher with GPS data regarding practice runs made by participants of the 2015 Ladies Run in Eindhoven and 2015 Marathon of Eindhoven. While these GPS tracks span all provinces of the Netherlands, they are assumed to be densest in the province of Noord-Brabant. To this end only the tracks in and around the city of Eindhoven are taken into account.

Important in selecting the research area is that the landscape in it is varied. One of the cities in Noord Brabant therefore seems like a logical option, as you have the cityscape and a less populated area around it to work with. An added benefit is the previously mentioned expected density of tracks in and around a city. As the largest city in the province and host of the two runs for which the GPS data was gathered, Eindhoven is the logical choice to research. Its municipality and the municipalities around it are selected as case study. These municipalities around Eindhoven are, starting to the north and turning clockwise: Son and Breugel, Nuenen, Gerwen en Nederwetten, Geldrop-Miero, Heeze-Leende, Valkenswaard, Waalre, Bergeijk, Eersel, Veldhoven, Oirschot and Best. Figure 1.1 shows the research area of this thesis. The chosen research area offers a great variety in landscapes, with ample bodies of water, natural areas, towns and cities, and agricultural land. The varied environment suggests also a varied spectrum of spatial influences in the province, which benefits the completeness of this research.

1.6 Research Scope

The scope of the research, i.e., what this research is about, is reflected in the research objectives, research questions and the case study. However, it is also important to elaborate on what will not be covered in this thesis. The first and most important part which also became apparent from the research objectives, is that reusable maps and tools are not a part of this master thesis. These more delivery focused results are not the main focus of this thesis, as that is on the formulation of a methodology. Other constraints on the scope are:

- The research area is a small part of the Dutch province of Noord-Brabant, the area in and around Eindhoven. While ideally the methodology can also be used elsewhere in the country, locations outside of the research area are not part of the scope.
- The research subjects are runners of all ages that participated in the 2015 Ladies Run and Marathon of Eindhoven. While results can hopefully be generalized to all runners, these are not explicitly taken into account.
- The influence factors only consist of things that can be realistically changed and manipulated. This means that for example the influence of temperature or broadly speaking, the climate, on runners is not researched in this thesis.
- The way in which the spatial influence factors are presented in the research methodology depends on the data that can be used for it. This means that the precision of the research results could be limited by the available data.
- This research utilizes only quantitative data, which means it is represented by numbers or attributes, during the analyses. While for comparison qualitative data (written or spoken text) would be of great added benefit to the analysis, this data is not available to the researcher.
- The methodology used in this thesis extensively utilizes ArcGIS and most importantly, its Python application ArcPy. This application is therefore not part of open source software and behind a financial barrier for most people. The methodology is, however, also applicable by using open source Python libraries, some of which are already used in the current methodology.

1.7 Thesis Structure

With the introduction to the topic and the setup for the thesis discussed in this chapter, the next chapter comprehends the information that can be deduced from related studies in a theoretical framework. The theoretical findings are thereafter operationalized in chapter 3, which is the outline of the research methodology. The research subjects, which were briefly discussed in paragraph 3.5, are explored extensively in chapter 4 for their characteristics through descriptive statistics.

The elaboration of the research methodology starts in chapter 5, wherein the spatial influence modeling of the influence factors is discussed. The enrichment of the research subjects with the spatial influences is the subject of chapter 6, wherein the methods to do so and initial descriptive results are discussed. The validation of these research results is explored in chapter 7, after which a conclusion to the master thesis and answering of the research questions is done in chapter 8.

Chapter 2
Theoretical Framework

2.1 Theoretical Research Challenges

In paragraph 1.2 the scientific relevance of this research has been discussed briefly. This paragraph expands on what kind of challenges can be encountered during this research from a theoretical point of view. These are identified by doing a literature exploration on the thesis subject among PABE literature.

To start, the research done by Ettema (2015) is analyzed, as it addresses some of those impediments. Firstly, he notes that research done on the influences of the built environment on runners is mostly qualitative in nature. This means that experiences and motivations of respondents are often discussed. Quantitative research on the subject or a combination with geographical data, however, has received little to no attention up until now. Quantitative research can be important in determining statistical assessments on how runners are influenced by the built environment and planning accordingly. In addition, doing analyses with geographical quantitative data can enhance the knowledge visually by presentation as maps, to make it easier to apply and deduce from. Since running is one of the most practiced outdoor physical activities, this lack in approaches is surprising (Breuer, et al., 2011).

Secondly, runners should not be seen as a uniform group when researching them (Ettema, 2015). In the PABE research done on runners, there is often a distinction made between different types or runners. Allen-Collison & Hockey (2007), for example, differentiated between joggers/fun runners, runners and athletes, a distinction that is made in most research. As can be expected, an athlete has very different needs and motivations to run somewhere than a jogger. When a practice round needs to be over 15 kilometer long an entirely different time and emotional investment is made than when the practice rounds are only 2 kilometer long. Groenink (2013) illustrated this distinction on a Dutch scale in a qualitative study. He found that joggers actually preferred a more lively urban setting on their runs, while more serious runners preferred the Amsterdamse Bos natural area, as it is more quiet and easy to run for longer distances uninterrupted. A quantitative look or the inclusion of geographical information, however, is still mostly absent in his research. Besides this distinction, runners can be distinguished in many more ways, for example based on age, gender or socio-economic status. Due to the focus on formulating a methodology, these distinctions are not considered during analysis.

Thirdly, as mentioned in paragraph 1.2, the influence of the built environment on hikers has been explored more extensively. It is, however, not sensible to assume that an environment fit for hikers is

also fit for runners. Ettema and Smajic (2015) provide a clear example on this issue. For their paper they interviewed over thirty people on their experiences with what kind of built environment stimulates them. An opinion that kept coming back was that they preferred lively environments e.g. markets, shopping streets or cafes. In contrast, it is safe to assume that runners do not benefit from lively areas with lots of people, as these provide hindrances on the route (Reiling & Dolders, 2015). Other stimuli or hindrances experienced by walkers could just as well have that same effect on runners:

"…in contrast to the presence of people engaging in leisure, the presence of too many other people engaging in transportation may be experienced as stressful or dangerous."

"…a lack of stimuli may be experienced as positive if the environment contains natural elements, such as trees and water…" (Ettema & Smajic, 2015, p. 108).

These two arguments are in line with scientific findings considering runners (Allen-Collinson, 2008; Bodin & Hartig, 2003). What these examples illustrate is that when the factors from the built environment are chosen in the next paragraph, literature about hikers can help determine the factors, but great care must be given when doing so. This way walkability literature is mostly used as backup information for this research, not the primary source.

A final theoretical impediment that can be identified is the simple fact that running does not take place in a designated arena. All public space can be used by runners to practice their activity. This can make it more difficult to find consistent influences in the built environment that stimulate activeness in runners or hinder it (see paragraph 2.4). This is especially evident in the evolution of running in the perception of the public, meaning the change from doing it predominantly on athletic tracks to doing it outside of these designated arenas (Scheerder & Breedveld, 2015). Shipway and Holloway (2010) further this argument by pointing out the increased popularity of distance running. The major drive behind this development is the accessibility of running events nowadays for both athletes and normal people. When more and more people run for great distances e.g. more than ten kilometers instead of small ones, however, stimulating it becomes more complex, specifically in a dense country like the Netherlands. When running these great distances, runners are more likely to encounter different environments and also hindrances in general. This leads in turn to a greater chance of demotivating the runner. Connectivity and ample supply of stimulating environments therefore seem necessary for successfully stimulating runners. Discussion on this problem, however, has also been lacking in PABE research.

2.2 Spatial Influence Factors

When considering the mostly qualitative research, the inherent differences within runners as a group, the inability to compare hikers and runners without consideration and the unpredictable performance area of a runner, the actual factors in the built environment that influence runners either positively or negatively can be explored, as all athletic performance is influenced by environmental factors (El Helou, et al., 2012). The paper *'Runnable Cities: How Does the Running Environment Influence Perceived Attractiveness, Restorativeness and Running Frequency'* from Ettema (2015) is used as the foundation of this exploration. Besides this article, an observation made by Allen-Collinson (2008) is also very important when determining the factors:

"Maintaining momentum, enhancing performance, and avoiding injury are the main qualities that a running environment should offer." (Allen-Collinson, 2008, p. 46)

This quote from Allen-Collison (2008, p.46) is used in conjunction with a paper from Ettema (2015) as the basis of the theoretical research into the influence factors. The first as the categorization of what each factor must uphold to and the latter as the starting point of expanding the theoretical scope. As mentioned before, while research about hikers needs to be handled with care, it is used to shed additional light on the found factors or provide contrasting views.

A final primary source of information used is a master thesis done by Reiling and Dolders (2015). As mentioned in paragraph 1.2, their thesis focused on runner friendly city design and not the formulation of a methodology and integrating geographical data in the research. Important to note, however, is that for the determination of factors, looking at the results of their thesis and which sources they used as a theoretical background is useful.

The factors have been subdivided into five different categories, which are discussed in the following order; running surface, social safety, traffic safety, surrounding environment and route information. At the end of the paragraph, all factors and their influence have been summarized in table 2.1.

2.2.1 Running Surface

As said, the factors regarding the running surface are discussed first. A maybe obvious explanation for this category is that the surface on which the activity is performed is an important determinant for runners to run or not run a certain route again. When looking at the observation of Allen-Collinson (2008) above, this category is directly involved with avoiding injury. Ettema (2015) makes the distinction between smooth surfaces, such as grass or pavement, and uneven surfaces, such as muddy roads. This distinction is made because uneven terrain is associated with an increased risk of injuries. These observations are mostly applicable to runners that are looking for short or interval running. When runners really want to build strength and endurance, running on uneven or otherwise difficult terrain is seen as beneficial (Hockey & Allen-Collinson, 2006). Another study by those authors uses the same distinction as Ettema, arguing that a smooth surface, such as grass, is key in avoiding degradation of the joints (Allen-Collinson & Hockey, 2013). There is, however a second distinction that is important to note, which is that between soft and hard surfaces. For example, for runners living in a city, running on the sidewalks is a risk factor for injuries. Harder surfaces are more dangerous in this case because they provide greater shock for a runner's joints to absorb. In contrast, running on surfaces where there is a lot of hindrance on them also results in a greater injury risk among runners (Johnston, e.a., 2003). To this end, five different factors about the running surface are taken into account: smooth, uneven, hard, soft and hindrances.

2.2.2 Social Safety

Besides running surface, four other categories can be made to house influential factors, two of which are safety related. The first of those two, social safety, is discussed first. Social safety contains factors

about the interaction between runners and their surroundings that make them feel safe. The first of which is how well lit the routes runners use are. This factor is obviously important when there is no natural light present e.g. evening, night and early morning. Despite this time restriction, it is often one of the most experienced impediments in PABE research (see Ettema, 2015). When the street lighting is poor, it is associated with an increased chance of injury (Allen-Collinson, 2008). Addy et al. (2004) found that when the street lighting is good, it stimulates people to be physically active more often. This conclusion is mirrored in the research done by Lee and Moudon (2008), where the research group even identified it as the most important facilitator of physical activity. Other research, however did not find a significant effect of street lighting on running behavior, while it was often indicated as an influential factor by respondents. Poor street lighting could even be associated with more quiet areas, and therefore have a surprising positive effect on physical activity (Ettema, 2015). This reversed effect will be explored in more detail in section 2.2.4.

Besides the influence of a physical factor on social safety, there are also social aspects that have an influence on it. Defined as a factor that significantly influences running frequency by Ettema (2015) is verbal harassment. This means that runners experience that people they come across say negative things about the runner in a way the runner can hear it. While it was found that it was not experienced often as an impediment, in contrast to the frequent experience of poor street lighting, it did have a significant impact. Runners experiencing verbal harassment were found to run less frequent compared to the runners who did not experience the phenomenon. Addy et al. (2004) came to a similar conclusion by defining it as one of the important barriers to physical activity. Clark (2015) even takes it a step further by suggesting that women might even experience more than only verbal harassment when running outdoors:

"The effects of harassment and perceived threats may be particularly relevant to girls or women whose activities demand a wide reign of outdoor space." (Clark, 2015, p. 1018)

While Clark focusses on girls and women in her paper, it is no leap to assume that the same impediment is perceived by boys and men when confronted with it, though to a lesser extent. A way to measure this is proposed by Addy et al. (2004). By looking at criminality rates for neighborhoods it can be estimated if harassment of any kind could be a barrier when running there.

The third factor that can be drawn from the social safety is the interaction with dogs. It is not uncommon for unleashed dogs to provide a hindrance to runners (Hockey & Allen-Collinson, 2006). Ettema (2015), while first stating that dogs probably have a negative influence on the attractiveness of certain running routes, found that hindrance from dogs was positively associated with attractiveness in his research. A possible reason for this is that the areas where people walk their dogs are probably quite similar to the areas runners like to run in. Noted in the paper, however, that this conclusion on the effect of dogs on runners is not conclusive and more research was needed. Allen-Collinson and Hockey (2013) in turn did find dogs as a contribution to hazardous places. They suggested that runners aim to avoid places which are popular for walking dogs.

2.2.3 Traffic Safety

The second category associated to safety is that of traffic safety. The theme from the Allen-Collinson quote at the start of this paragraph that it is best associated with is the 'maintaining momentum'. The most important limitation for runners in maintaining their momentum is that they have to stop for other traffic on occasion. This is especially relevant for urban areas, where the traffic networks are denser and more utilized than in the countryside (Reiling & Dolders, 2015). In this category a distinction is made between motorized vehicles as a nuisance and cyclists as a nuisance. The second category is especially important in the Dutch context.

The presence of motorized vehicles is probably the most obvious category of the two. Runners often have to perform their activity in close proximity to them, increasing the danger, or have to cross a busy road. Ettema (2015) found that having to stop for traffic negatively affects the attractiveness of a running route, meaning the runner is less likely to run there in part because of traffic hindrances. Boarnet et al. (2011) also found that the need to stop for other traffic negatively impacts walkers. To mitigate this influence, clear places to cross roads, sidewalks and other indicators to increase traffic safety are proposed. In this case, however, the influence on walkers and runners might differentiate, not fundamentally, but ever so slightly. In the case of runners, losing momentum seems more severe than for walkers, making this factor one of the most important barriers to physical activity (Addy et al., 2004; Jongegeel-Grimen, et al., 2013).

As said, avoiding cyclists while running might be a factor that is of specific importance to studies focusing on the Netherlands. With the abundance of cyclists in Dutch cities, they can easily get in the way of runners and vice versa (Reiling & Dolders, 2015). While in the urban environment, it might be difficult to avoid them altogether, it is important for runners both inside and outside of cities to avoid this hindrance as it disturbs the momentum of running and could even cause injury when collisions happen (Allen-Collinson, 2008). While cyclists are often coupled as a nuisance to runners with pedestrians, the latter are excluded from this research with relation to how these factors are operationalized with data. Ettema (2015) implies that the effect cyclists have on runners could be related to their personal safety, of which the traffic safety is a segment.

2.2.4 Surrounding Environment

In this section some of the most important spatial influences from the literature considering the environment to run through are discussed. These factors can be associated with the 'enhancing performance' part of the quote by Allen-Collinson (2008). As a start, perhaps the most obvious and important stimulating environment, which has been mentioned occasionally in the previous sections, is discussed. This is the presence of natural or 'green' areas. Bodin & Hartig (2003) researched if the effect of natural areas on runners is stimulating in the same way as it is on hikers. They came to the conclusion that runners indeed preferred being in green environments over the urban environment, as it has a more restorative effect on them. Gladwell et al. (2013) built on this perception and suggest that considering the role of nature as a tool is essential when mitigating physical inactivity. Reasons for this suggestion are that it stimulates revitalization and even self-esteem, while decreasing tension, anger and depression. Perhaps most importantly, however, is that they found that green areas also have a

positive effect on the perception of effort by runners. This in turn can lead to increases in motivation and the intensity of physical exercise. These results are mirrored by other research, confirming the importance of natural areas in stimulating outdoor physical activity such as running (Shipway & Holloway, 2010; Loureiro & Veloso, 2014; Ettema, 2015; Reiling & Dolders, 2015). Important to note is that not only the green environment is part of this influence factor, but also the blue one. The presence of water is also seen as a natural stimulus for runners. The body of water could be a sea, lake, river or small creek, its influence has mostly been documented as positive (Ettema & Smajic, 2015).

A factor that could be seen as closely related to natural areas is the relative silence of an area and the influence it has on runners. Reiling and Dolders (2015) in their master thesis suggested that silence was an attribute inherent to green areas. While it is safe to assume these areas on average are less noisy than urban areas, it might be a wrong assumption that silence is completely linked to these areas. As mentioned in section 2.2.2 on social safety, the absence of street lighting could also be an indicator of the noisiness of an area (Ettema, 2015). When looking at hikers, the results are contradicting. While they enjoy lively areas when walking around, too much stimuli can have a reverse effect. In that case, the quietness of an area can become a positive attribute (Ettema & Smajic, 2015). Comparing this to runners, however, it is safe to assume quieter areas in general will have a more positive influence on them than lively areas. With too much stimuli, the values of maintaining momentum and avoiding injury (Allen-Collinson, 2008) are endangered.

While natural areas itself are found to have a stimulating effect on runners, there is an important aspect they and other environments as well need to have for the positive influences to optimally manifest themselves, which is diversity. Only passing the same trees over and over gets boring, but when for example different kinds of natural areas are mixed or some elevation is present, it facilitates the positive influences on runners (Reiling & Dolders, 2015). Another form of variation important to runners' motivation is a variation of function. This means that the built environment becomes more stimulating when different land uses are passed somewhat frequent by a runner. An example could be to first pass an office area, then a city park, thereafter running parallel to a canal and ending with a suburban residential quarter. When the aesthetics change regularly it has a positive influence on the runner (Lee & Moudon, 2008; Ettema, 2015).

A fourth factor in this category is the influence of air quality on the runner. This influence came to light when the 1984 Olympics were to be held in smoggy Los Angeles and again leading up to the 2008 Beijing Olympics. It was found that bad air quality indeed affects athletic performance, and thus also running. The major influence it has on runners is that the oxygen intake is impaired by the oxidants (smog) and reductants in the air (McKenzie & Boulet, 2008). When running in polluted air, the respiratory system has an increased workload, resulting in a faster drain of the runner's energy Shephard, 1984). This effect seems more influential the longer a physical activity takes. A positional based solution mitigating this influence is to avoid industrial areas and roads with a high traffic volume. This factor therefore relates both to the importance of 'enhancing performance' and 'avoiding injury' (Allen-Collinson, 2008). Falt (2006, p. 268) adds to this by stating:

"… in some communities in California where air quality is poor, the most athletic children are three times more likely to suffer from asthma than their peers who do not exercise."

The final fifth factor in this category is the sum of all other factors, as it were. It is argued that if people live close to stimulating environments they are more likely to be stimulated by them and participate in

physical activity (Ettema, 2015). In addition to runners, hikers also benefit from convenient or nearby opportunities to perform their activity (Addy, et al., 2004; Lee & Moudon, 2008).

2.2.5 Route Information

In this section the effect of available route information on runners is explored. When properly informed about where ´good´ running routes are or how far they still need to go, runners are more likely to run these routes (Ettema, 2015; Reiling & Dolders, 2015). A good example of this is a route around the Sloterplas in Amsterdam that is just shy of six kilometers. At regular intervals, the distance already covered is put on the ground, as can be seen in figure 2.1. This results in runners being able to track their progression and performance.

Figure 2.1: Route information next to the Sloterplas in Amsterdam (van Poortvliet, n.d.)

Factor	Influence	Walkability source	Runnability source
Running surface			
Smooth	+		Hockey & Allen-Collinson, 2006 (+); Allen-Collinson, 2008 (+); Ettema, 2015 (+); Allen-Collinson & Hockey, 2013 (+)
Uneven	-		Hockey & Allen-Collinson, 2006 (+/-); Allen-Collinson, 2008 (+); Ettema, 2015 (+)
Hard	-		Johnston e.a., 2003 (-); Allen-Collinson & Hockey, 2013 (+/-)
Soft	+		Johnston e.a., 2003 (+)

Hindrances	-		Hockey & Allen-Collinson, 2006 (-); Johnston e.a., 2003 (-)
Social safety			
Street lighting	+	Addy, et al., 2004 (+); Lee & Moudon, 2008 (+)	Allen-Collinson, 2008 (+); Ettema, 2015 (+/-)
Verbal harassment	-	Addy, et al., 2004 (-)	Ettema, 2015 (-); Clark, 2015 (-)
Dogs	-		Hockey & Allen-Collinson, 2006 (-); Allen-Collinson & Hockey (-); Ettema, 2015 (+)
Traffic safety			
Motorized vehicles	-	Addy, et al., 2004 (-); Boarnet, et al., 2011 (-)	Hockey & Allen-Collinson, 2006 (-); Jongegeel-Grimen, et al., 2013 (-); Ettema, 2015 (-)
Cyclists	-		Allen-Collinson, 2008 (-); Reiling & Dolders, 2015 (-)
Surrounding environment			
Natural areas	+	Loureiro & Veloso, 2014 (+); Ettema & Smajic, 2015 (+)	Bodin & Hartig, 2003 (+); Shipway & Holloway, 2010 (+); Gladwell et al., 2013 (+); Ettema, 2015 (+); Reiling & Dolders, 2015 (+)
Silence	+	Ettema & Smajic, 2015 (+/-)	Groenink, 2013 (+/-); Ettema, 2015 (+); Reiling & Dolders, 2015 (+)
Varied environments	+	Lee & Moudon, 2008 (+)	Allen-Collinson & Hockey, 2013 (+); Ettema, 2015 (+); Reiling & Dolders, 2015 (+)
Clean air	+		Falt, 2006 (+); McKenzie & Boulet, 2008 (+)
Closeness to stimulating environments	+	Addy, et al., 2004 (+); Lee & Moudon, 2008 (+)	Ettema, 2015 (+)
Route information			
Route information	+		Ettema, 2015 (+); Reiling & Dolders, 2015 (+)

Table 2.1: Factors in the built environment influencing runners

2.3 Geographic Information Modeling Problems

Table 2.1 shows the diverse range of spatial influences that are and can be determined when using the PABE scientific literature. As mentioned before, one of the main contributions of this study to that field is the implementation of geographical information when discussing runners. Therefore, it is also essential to discuss the problems that arise when dealing with geographical information.

"Any study that examines the effects of area-based attributes on individual behaviors or outcomes faces two fundamental methodological problems." (Kwan, 2012, p. 958)

With the above mentioned quote, Kwan (2012) opens her paper on the 'uncertain geographical context problem' [UGCoP], one of the two problems hinted at in the quote. The other problem hinted at is in her opinion more well-known and given more attention, but is not more important. This is the 'modifiable areal unit problem' [MAUP]. In this paragraph, both are discussed together with their relevance to this research.

2.3.1 Modifiable Areal Unit Problem

This problem was originally thought to be exclusive to the human geography field of science. Over time, however, it became clear that this was not the case and that it was relevant to any study working with remote sensing or GIS. As long as areal units are used in the data for research a scientist must be careful about the effects of the MAUP and how to deal with it (Dark & Bram, 2007). In the SAGE Handbook of Spatial Analysis, Wong (2009) defines the essence of the MAUP as:

"…there are many ways to draw boundaries to demarcate space into discrete units…" (Wong, 2009, p. 106)

The complete description used in the handbook is focused on the field in which the problem is most well-known, namely administrative boundaries. While this research focusses on other subject fields, the meaning of it stays the same. When looking at different datasets describing a common phenomenon, it is quite possible the spatial areas in those datasets are not the same. This can be due to human error when gathering and compiling the data, but it can also be the result of the data being snapshots of different times between which boundary shifts have taken place. The effects of the MAUP can cause statistical errors when working with the data (Dark & Bram, 2007). Menon (2012) suggests two ways in which the distortions or errors can be minimized:

1. The areal units of analysis are identical in terms of shape, size and neighboring structure.
2. The areal units of analysis are spatially independent.

As argued thereafter, however, meeting these prerequisites is difficult, meaning that having an understanding of the MAUP and methods to deal with it is important when working with data containing areal units, be it vector or raster data (Menon, 2012).

There are two main effects on research results that signify a MAUP; the scale effect and the zoning effect (Openshaw & Taylor, 1979). The scale effect indicates that analyzing the same data, but on different scales can lead to vastly different results. When looking at a phenomenon on a national scale,

a different result is to be expected than when looking on a municipality scale. The ground rule here is that the scale cannot be smaller than the scale of the data itself. The zoning effect is a problem that could still happen even though all data is looked at on the same scale. It indicates that based on the division made within the data objects, the results are variable. This particular effect could also be used to manipulate results. The most obvious example of this is gerrymandering (Wong, 2009).

The question that needs to be asked is how the MAUP is relevant to this research and to what extent. To answer it, the MAUP is relevant to any research working with geographical data, meaning this thesis too. Considering, however, that the primary focus is on formulating a new methodology and not on results, it might not be the most important problem to face of the two explored in this paragraph. When looking at the factors that are proposed to be included in the analysis, and therefore need to be modelled, multiple sources of MAUP errors in results can already be expected to occur. Most obvious is that some factors are going to be modelled by means of different datasets. When these datasets do not utilize the same regions, it needs to be checked how this influences the integrity of the analyses, in other words how big the influence of the MAUP is for the influence factor.

2.3.2 Uncertain Geographical Context Problem

The second problem Kwan (2012) mentions in the quote at the start of this paragraph is the UGCoP. This problem exists because the geographical contexts used in geography research when working with the effects of area-based attributes on individuals is a complex feature to determine and is often done in different ways. There is no *"true causally relevant"* geographical context determined in geographical data to use as of yet, because of these complexities. The question is often whether the given geographical context is likely to be reasonable proxy for the true causally relevant geographical context (Diez-Roux & Mair, 2010). The fact that the geographical context is modeled and delineated in different ways in research is probably one of the reasons why the results of effects of social and physical environments on health behaviors and outcomes are often inconsistent (Kwan, 2012). This can even be the case for different explorations of the same phenomenon in the same research area. Two sources that contribute to the geographical context uncertainty can be identified:

1. The uncertainty in the spatial configuration of the appropriate contextual units.
2. The uncertainty about the timing and duration to which individuals are exposed to the contextual influences (Kwan, 2012, p. 959).

There are many different spatial configurations used for contextual units in geography research to date. In general they can be put into either of three categories. The first and most used type of spatial configuration is to use administrative boundaries. Data often is tied to administrative boundaries, making this not only a convenient but also sometimes the only viable option. However, using these kinds of boundaries present some problems. Simply assuming that the correct geographical context is the boundary of a neighborhood is often an oversimplified method. Only when the influence indicators refer exactly to an administrative boundary as the geographical context, this method can be seen as a reasonable proxy of the actual geographical context, a scenario that is difficult to achieve (Chaix, 2009). When using administrative boundaries of places of living as the geographical context, another problem arises. For most people, even daily life takes them to other neighborhoods and cities, be it for work, education or even something minute such as grocery shopping (Kwan, 2012).

The second category of spatial configurations used to determine a geographical context is to make use of the perceived neighborhood of the research participants. With this method, social interactions and routines and people's perceptions of their neighborhood need to be expressed spatially, which leads to a geographical context. It is quite probable this context is wildly different from any possible choice of administrative boundary. Problem with this approach, however, is that in reality parts of the participant's daily activities would still happen outside of their perceived neighborhoods, which leads to the used geographical context being a less likely proxy to the true geographical context (Kwan, 2012). A great example of this is the result from the research of Basta, Richmond and Wiebe (2010), as shown in figure 2.3. Ten subjects got to draw their perceived neighborhood, after which their movement were tracked using Global Positioning System [GPS]. Of the ten people, only one actually stayed in the area he or she determined as the neighborhood.

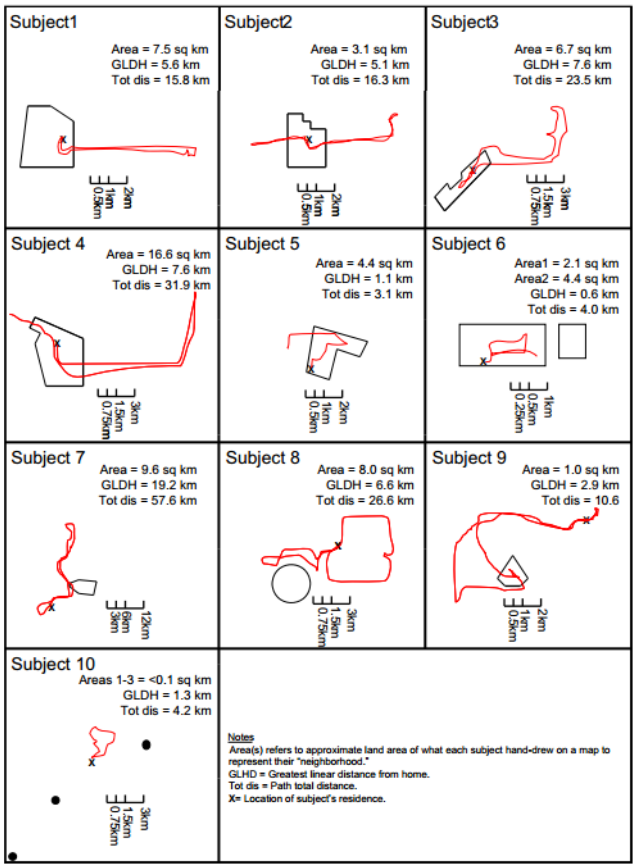


Figure 2.2: Paths of 10 subjects' daily activities of one day, in relation to their perceived neighborhood and residence (Basta, Richmond & Wiebe, 2010, p. 1947)

The third approach to spatial configuration of contextual units is to use a fixed area around the participant's home. Kwan (2012) gives the examples of a circular zone around the home, a certain distance road network buffer or a certain distance or time walk radius. She argues, however, that it is far from clear how well these methods approximate the true geographical context. Besides these spatial configurations the geographical context is also dependent on the features of the research participants, e.g. when researching elderly it is safe to assume a smaller context is needed than when researching

runners, as is done in this thesis. To close the discussion, the first source of geographical context uncertainty is summarized in a deduction made by Kwan (2012, p.960) from the book 'Mobilities and health' by Gatrell. It is relevant in painting the conundrum faced when deciding on a spatial configuration for the geographical context:

"The multilevel and multiscale nature of contextual influences greatly complicates the task of accurately delineating the appropriate contextual units, which could be nested or overlapped in a complex manner. Part of the uncertainty in the spatial configuration and boundaries of contextual units arises from the dynamic characteristics of individuals and contextual influences."

As for the manner of the second uncertainty Kwan (2012, p. 959) stated, that of unknown timing and duration of exposure by the contextual influences on individuals, it is a problem that is mapped increasingly well due to new positional techniques such as GPS. Recent research found that people spent a lot of time outside the conventional static spatial configurations used for a neighborhood. Not only that, but there also seems to be a great variability in how much people spent inside the static context from day to day. This leads to an unclear picture of how much contextual influences people actually get when using a static context (Wiehe, et al., 2008; Basta, Richmond & Wiebe, 2010). When armed with such knowledge it becomes clear that not only the spatial, but also the temporal configuration of the geographical context can be quite problematic. There is no universal timeframe to research certain phenomena. It is dependent on the research participants, the phenomena to be explored and even things such as culture and country in which the research is held. A final facilitator of the uncertainty of timing and duration of contextual influences is that a data collection done of time t often is the result of contextual influences of an unknown period before t (Kwan, 2012). Geographical research that does not address the UGCoP often makes the assumption, be it knowingly or not, that the participants' life through space and through time stays the same during the period in which they are observed.

The two contextual uncertainties are also of importance when dealing with GPS measurements, such is the case in this thesis. While the examples used in this paragraph are mostly related to places of living of people, a GPS measurement is a position in space in the same way. When looking at a measurement and an influence of a spatial factor, the choice still needs to be made how the geographical context is modelled. Are administrative boundaries the way to go, are respondents themselves asked how they would see the geographical context around the measurement or is drawing a certain area around the measurements the best approach? As the researcher is not in contact with any of the respondents, letting them decide the geographical context is not an option, but the other two are viable. With inclusion of temporal information in the GPS data, the timing of the measurement and duration of an entire run is also relevant to this research. An obvious example is the importance of street lighting as a spatial factor in the evening and night over its importance during the day.

Considering the inclusion of multiple factors (see paragraph 2.2) a correct proxy of the geographical context for each of them needs to be determined, leading to overlapping contexts. Modeling a good proxy for each factor, however, is of vital importance for the viability of the methodology and can also be a good contribution to PABE or GIS research on its own.

2.4 Conceptual Model of Factors

To get a good overview and scope of the theoretical background discussed in this chapter, this paragraph entails a conceptual model. Figure 2.2 shows the model, in which the influence of factors on runners is schematically shown. As can be seen, a runner performs a run on route A during which he experiences the different spatial influences, as discussed in paragraph 2.2. These give a sum that encompasses the complete set of positive and negative influence the runner experienced on Route A. As is to be expected, this differs from route to route. The influences are shaped by their geographical context. As discussed in paragraph 2.3, the MAUP and UGCoP are to be considered while determining the context of each spatial influence.

Together, this forms the theoretical scope of this thesis. The conceptual model also shows a decision that is made after each run. Will route A be used again by the runner, or will a different option (route B) be used when opting to run again. The decision making process of physical activities is often part of the focus of contemporary PABE research (Ettema, 2015). However, as the psychological part of spatial influences on runners is not part of the research scope, this step is not discussed in the remainder of this thesis. In the next chapter, the operationalization of this scope is discussed.

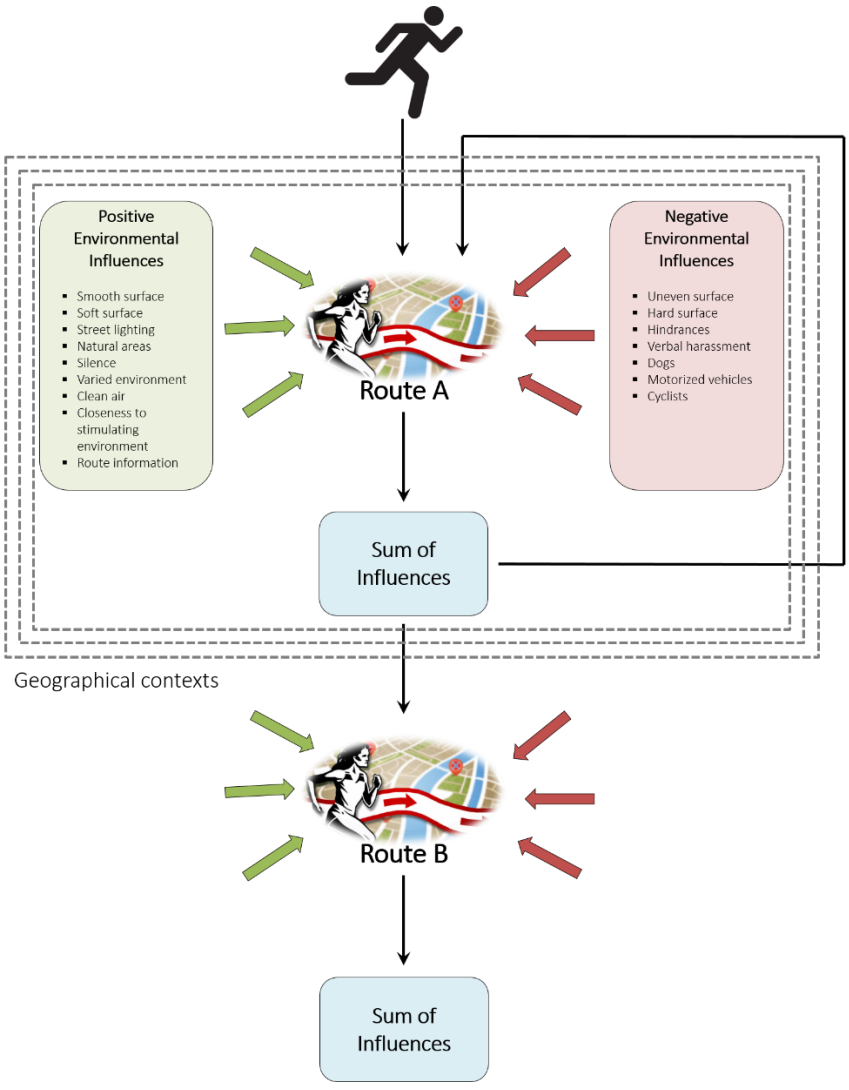


Figure 2.3: Conceptual model

Chapter 3
Operationalization

3.1 Introduction

The previous chapter was about gathering the necessary theoretical information to back up the formulation of the research methodology. This chapter goes into detail how the conceptual model shown in figure 2.2 can be operationalized into a methodology. The chapter is split into different paragraphs that mostly follow the structure that is the resulting workflow. The methods in which a spatial influence could be modeled is discussed first, after which the used software and data, the modeling of the spatial influence factors, the enrichment of route data with those influences and the validation of the preliminary results are discussed. As mentioned, the result of this chapter is a workflow of the whole research methodology and is discussed in paragraph 3.6.

3.2 Modeling the Influence Functions

With the possible influence factors on runners discussed, the next step is determining ways in which these and their spatial context can be modeled. The MAUP and UGCoP have already been discussed in paragraph 2.4. The complications these two problems pose require multiple possible methods to model a geographical context. The way the context is modeled and the parameters used means their values are bound to differ from factor to factor. Methods to model the geographic context are split into three categories: surface based, distance based and area based. These categories are discussed in this paragraph.

3.2.1 Surface based influence component

The first way to model the geographical context is to use a surface based influence component. Before explaining this category, it is important to note that each factor has an influence value [V], which does not change and thus is static, and an influence probability [P], which is variable and signifies the probability of an influence source influencing a runner with value V.

The surface based influence component is the most straightforward option of the three. It is only applicable to influence factors that can be aggregated to the street network used during analysis. The principle is that the spatial influence of a factor is only relevant if the runner is physically on a running surface in the network. The values for the different running surfaces differ from surface to surface (see paragraph 5.1). The probability for surface influence components is quite static, however, despite what is previously mentioned about it at the start of this section. If a runner is on a surface there is a probability of 1 that the runner experiences the spatial influence. If the runner is not on a running surface that is part of the network, the P is 0 and no spatial influence is experienced. This component is visualized in figure 3.1.

GPS measurements have an inherent positional error. Depending on the device used for the measurements, the surroundings of the GPS positions and the method in which the signal from the device is matched to the signal of a satellite, this can differ from a couple of centimeters to tens of meters (Bona, 2000; Newson & Krumm, 2009). For example, both real and urban forests can lead to noise for GPS devices that hurt the precision and accuracy of the measurement, because of their vertical nature. Yoshimura and Hasegawa (2003) found that horizontal positioning errors were relatively small when using roads in forested areas compared to going off-road. Therefore, for this research, it is assumed that runners run on a running surface that is part of the street network, which should limit at least the horizontal positioning errors (Yoshimura & Hasegawa, 2003). A concept called map-matching is used to snap the measurements to a network segment. Map-matching is discussed in more detail later in this chapter in paragraph 3.4. The problem that can arise, however, is that a runner does decide to run off-road. How this problem is dealt with is discussed with the map-matching, but when this happens it does mean that spatial influences modeled with the surface based component are not experienced by the runner with this methodology. Additional modeling of these influences might be necessary in these cases, depending on the occurrence of it.

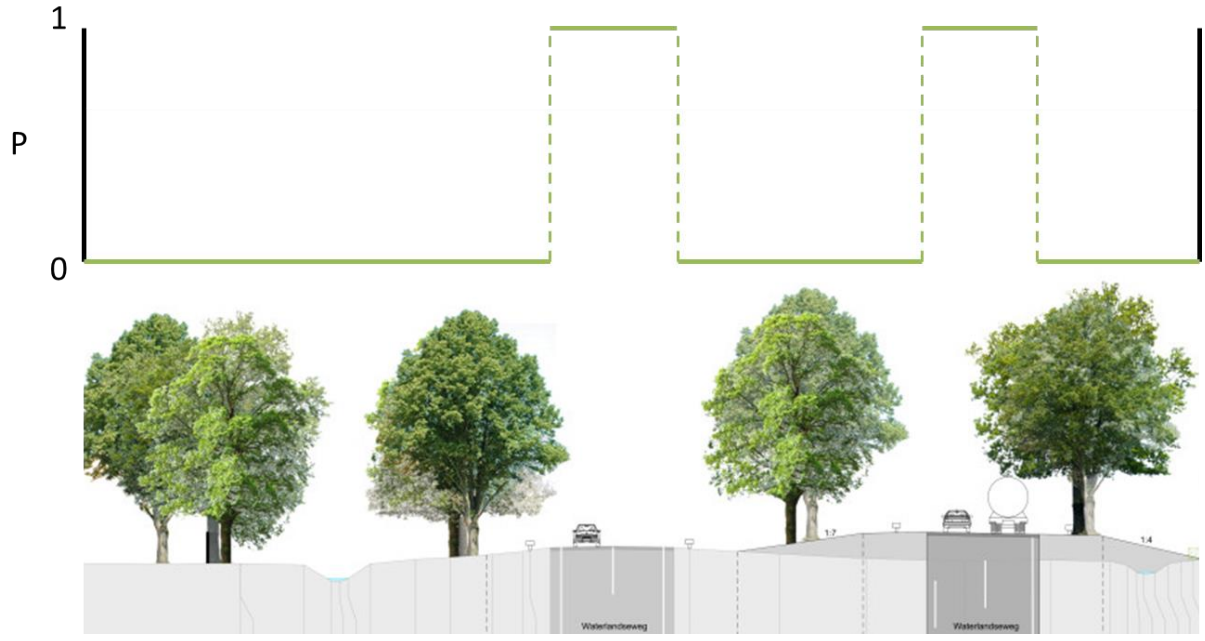


Figure 3.1: Boolean distance based influence example (Bureau-Maris, n.d.)

3.2.2 Distance Based Influence Component

The second method to model the spatial influences on runners is to use a distance based influence component. This means that influence sources closer to the runner have a greater probability of influencing the runner than sources further away. In figure 3.2 an example of this method is given with the probability of influence of the groups of trees on the car that drives on the road. In the example the function corresponding with this method for the P is:

(1) $$P(x) = \alpha^{-bx}$$

P = the influence probability of the spatial influence factor
α = a value that represents the value when the distance is 0
b = a value that represents the steepness of the function and thus the rapidness of the decline
x = an exponent that makes this function an exponent function and also represents the distance between influence source and the runner

In the example of figure 3.2, the closest group of trees (A) has a P of around 0,4, which means that there is a 40% chance that these trees have an influence on the car in the picture. The furthest group of trees (C) has a P of 0,05, which means that there is a 5% chance these trees have an influence on the car. This version of the distance based influence function makes use of exponential decay, which means the initial probability decline is steep and flattens out on greater distances, edging closer to 0.

Besides exponential decay as a distance component, linear decay is also possible to model the spatial influence with. Instead of an irregular decline of P over distance, this decline is regular. Figure 3.3 shows the linear decay function. Some added variables that can be determined are also shown in the figure. The first is the distance at which the decline starts, presented by F. For example, running 10 meters from a waterway might still give the same probability of influence as running right next to it. The other consideration is at which point the probability chance P becomes 0, in other words the steepness of the influence function. The linear decay influence function can be seen in the equation below. When a function is constructed wherein the decline does not start from a distance of 0, the b value is the value at 0 if the linear function is drawn all the way to 0, otherwise B will always have a value of 1.

(2) $$P(x) = aX + b$$

P = the influence probability of the spatial influence factor
aX = the steepness component
b = the value when the distance is 0

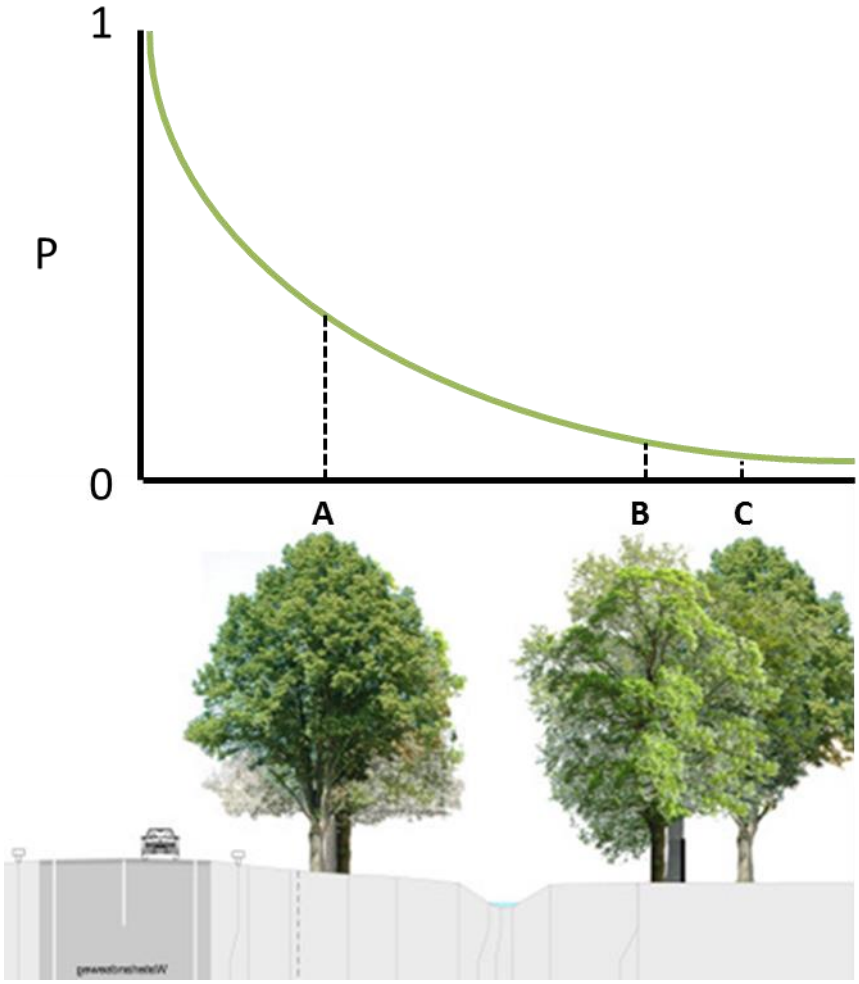


Figure 3.2: Distance-based influence function example 1 (Bureau-Maris, n.d.)

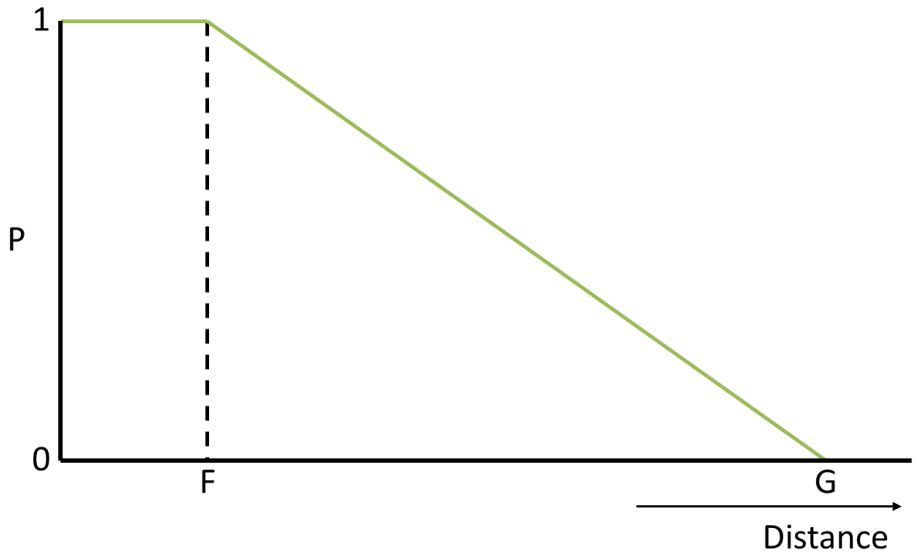


Figure 3.3: Distance-based influence function example 2

One important note to make concerning the distance decay method is that due to the horizontal positional error the GPS measurements are prone to, a buffer is often chosen around the influence sources when modeling the functions. This buffer is set to 10 meters and will mean that if a runner is within 10 meters of an influence source, the influence probability is set to 1, to counter the possible error in location (represented by F in figure 3.3). The probability decay therefore starts from 10 meter from the source. Exceptions are made when the influence decay too steep, meaning the buffer leads to more noise in the results than the positional error would.

3.2.3 Area based influence component

The third method proposed in this thesis to model the geographical context of an influence factor is to use an area based influence component. This method corresponds with the method proposed by Kwan (2012) to use a certain area around the research object as a geographical context (see paragraph 2.3). For the area based influence component an area around a GPS measurement is determined for a factor. All influence sources inside the area are then used to calculate the influence probability. With this method, a runner at the location of the GPS measurement can have a certainty of influence by an influence factor, despite none of the influence sources having a P of 1. It must be said that the influence probability cannot exceed 1, as that is a certainty of influence. This method can be summarized in the following equation:

(3) $$\sum(P * V)$$

P = the probability of influence happening from an influence source
V = a value assigned to different kinds of influence sources in a influence factor and can be based on a variable of the used dataset

A second method of using an area based influence component is to not use all influence sources inside the area, but only the one with the highest P. The equation below shows this method. When only using the max P, it must be considered in what way these probabilities are determined. Two proposed ways of doing that is to use either the Euclidean distance from the GPS measurement or to let the size of the influence source determine it. For example, an entire forest at 50 meter from the runner is likely to have a bigger effect on him/her than a couple of trees at 50 meter.

(4) $$\max P * V$$

In chapter 5, which spatial influence function is used for which influence factor is discussed extensively.

3.3 Software and Data

The two main components of any geographical analysis are the geographical data itself and the tools (GIS) used to analyze it (Goodchild, 2010). These two components are discussed in this paragraph, starting with the latter.

3.3.1 Used Software

The used software can be subdivided into two categories in this thesis. The first is software used for data manipulation and the second is software used for data visualization. While GIScience studies often also utilize GIS for data storage, no specific tool is used for that in this thesis, as this is done concurrently with analyses.

Manipulation and Analysis

The main tool that is used for data manipulation and analysis is the programming language Python, version 2.7. Python is an open source programming language that relies on first and third party libraries of functions. A lot of libraries, or modules, focus on manipulating geographical data, which, together with the programmer friendly orientation of the language (Python, n.d.), makes it useful for executing the preprocessing and analyses parts of this research. Another benefit of this is that a researcher can use Python as a way to connect other GIS, such as ArcGIS, FME and QGIS.

The second tool used for this purpose is FME, short for Feature Manipulation Engine. This GIS, engineered by Safe Software, is used sparingly when necessary. Only where Python is lacking, which is dealing with very big datasets and simple data type transformations, FME is used.

A final tool that is used for the data analyses is SPSS, a tool used for statistical analyses that can be applied to geographical data. While it is quite possible to perform most statistical analyses with Python, its benefit of interconnectivity with other software and the ease of use and availability to the researcher through Utrecht University make SPSS the more obvious choice for performing the statistical analyses.

Visualization

As said, data visualization is the other category of software used. The data visualization is done through either FME or ArcGIS. Visualizing the data is not a main focus of the research methodology or this thesis in general, but it serves the useful purposes of presenting the research results and checking the results mid-analysis. Because of the lack of focus on visualization, the two tools used for it are chosen based on previous experience and availability to the researcher.

3.3.2 Data Procurement

Another important limitation to take into account before assigning spatial influence methods to the influence factors is the available data that can be used. As there are many different influence factors determined in the previous section, different datasets need to be found to fit them. While some dataset can be used for different influence factors, some influence factors need multiple datasets to properly model.

The first part of this methodology phase is the procurement of data relevant to the spatial influence factors determined in Chapter 2. Analyses cannot be done on data that is not available to the researcher.

During the data procurement there is one limitation set upon the possible data, which is that it must be available to the researcher without any additional monetary costs. The University of Utrecht hosts a vast plethora of datasets that can be useful for geographic research, while other datasets are open source and available at other data portals. Table 3.1 provides an overview of the influence factors, the datasets utilized to model them and the source of each dataset.

As can be seen in table 3.1, there are some changes from how the influence factors were presented in table 2.1 and the availability of data. Firstly, the running surface is grouped into one influence factor. This is done as the same datasets are used for all the separate influence factors that belonged in this category, but different attributes belong to the different running surface aspects. The same grouping is not done to the motorized vehicles datasets, as it is expected they need to be modeled with different spatial influence modeling methods. Secondly, the influence factors 'dogs', 'closeness to stimulating environments' and 'route information' are excluded from further analysis, because no viable datasets are available to model them. In the case of the presence of dogs this is because municipalities have different ways of storing relevant data, if they even have it. Since twelve municipalities are used in this analysis, modeling this influence factor is not viable. The closeness to stimulating environments is excluded due to the lack of information on the runners themselves. As it is not known if they are starting from their houses or are travelling to the starting point of the route, it is impossible to model this influence factor. Route information as an influence factor is also excluded, as mentioned. This is due to the lack of available information and data.

Factor	Dataset(s) used	Data owner
Running surface		
Running surface	Top10NL_WEGDEEL_HARTLIJN	Kadaster
Social safety		
Street lighting	Fietsbondnetwerk_NL	Fietsersbond
Verbal harassment	Statistics Netherlands criminality rates	Statistics Netherlands
	Wijken en Buurten 2016	Statistics Netherlands
Dogs	-	-
Traffic safety		
Motorized vehicles	Top10NL_WEGDEEL_HARTLIJN	Kadaster
Cyclists	Fietsbondnetwerk_NL	Fietsersbond
	Wijken en Buurten 2016	Statistics Netherlands
Surrounding environment		
Natural areas	Top10NL_TERREIN_VLAK	Kadaster
	Top10NL_WATERDEEL_VLAK	Kadaster
	Top10NL_WATERDEEL_LIJN	Kadaster
Sound	cden16_k8	RIVM
Varied environments	Top10NL_TERREIN_VLAK	Kadaster
Clean air	PM10_Luchtmeetnet	RIVM

	PM2,5_Luchtmeetnet	RIVM
	NO2_Luchtmeetnet	RIVM
	EC_Luchtmeetnet	RIVM
Closeness to stimulating environments	-	-
Route information		
Route information	-	-
Additional datasets		
Runner tracks	EHV-clean.geojson	Technical University Eindhoven / Fontys Sporthogeschool Eindhoven
Gemeentekaart	Imergis_gemeentegrenzen_kustlijn	BrigGIS Geoservices BV

Table 3.1: Data sources used to model the spatial influence factors

- Fietsbondnetwerk_NL: This dataset contains a network dataset of every road a motorized vehicle or bicycles can legally travel on. In this dataset many relevant attributes are stored. Examples are how well a road is lit, the quality of the road, the amount of obstacles, the geometry and the traffic volume.
- Top10NL_WEGDEEL_HARTLIJN: Something the previous dataset lacks that is important for mapping runner routes are roads bicycles are not allowed on, but pedestrians are. This dataset from the Kadaster contains these roads, but is more lacking when it comes to attributes measured for the roads.
- Statistics Netherlands criminality rates: This dataset is gathered from the Statistics Netherlands and represents the criminality rates of postal code areas in the Netherlands of 2017.
- Pc4_single_vlak: This dataset containing polygons of the postal code areas in the Netherlands is used in conjunction with the previous one, to give geometry to the criminality rates. The dataset is from 2017, as the criminality rates are also from the same year.
- Top10NL_TERREIN_VLAK: This dataset part of the Top10NL database contains all areal functionalities except infrastructure and water. It does include all natural areas in the Netherlands.
- Top10NL_WATERDEEL_VLAK: This dataset contains all waterbodies in the Netherlands (lakes, seas, rivers wider than six meter, swimming pools, etc.). The dataset can therefore be used to model the 'blue' nature, but is not complete.
- Top10NL_WATERDEEL_LIJN: In addition to the previous dataset, this one contains all water streams up to a width of six meter. These two datasets completely cover the water bodies in the Netherlands.
- Cden16_k8: A dataset owned by the National Institute of Public Health and the Environment [RIVM] containing the information on sound pollution in the Netherlands. It combines sound pollution numbers from highways, other roads, railways, airports and airplanes, industry and wind turbines. This information is gathered in the period between 2011 and 2016. An example of this dataset projected on the research area can be seen in figure 3.4.

- Luchtmeetnet Datasets: In total four datasets to measure the air quality are used. Each represents a polluting air quality particle that is measured by different measurement stations around the country. These four are the datasets with measurements for bigger and smaller particles (PM10 and PM2.5), nitrogen dioxide (NO2) and soot (EC) (Atlas Leefomgeving, 2018a). An example of the air quality in the research area can be seen in figure 3.5.
- Gemeentekaart: As the research area is the Dutch city Eindhoven and the municipalities around it, a dataset needs to be used to get a shapefile of this area to cut the other datasets and speed up processing time. This dataset is from 2017 and is owned by BrigGIS Geoservices BV.
- Runner tracks of 2015 Ladies Run and Marathon of Eindhoven participants (EHV-clean.geojson): This is the dataset that is used in the analyses of this thesis. The dataset consists of practice GPS tracks from participants of the 2015 Ladies Run and Marathon of Eindhoven, as mentioned before. In chapter 4, the descriptive statistics of this dataset are explored in greater detail.

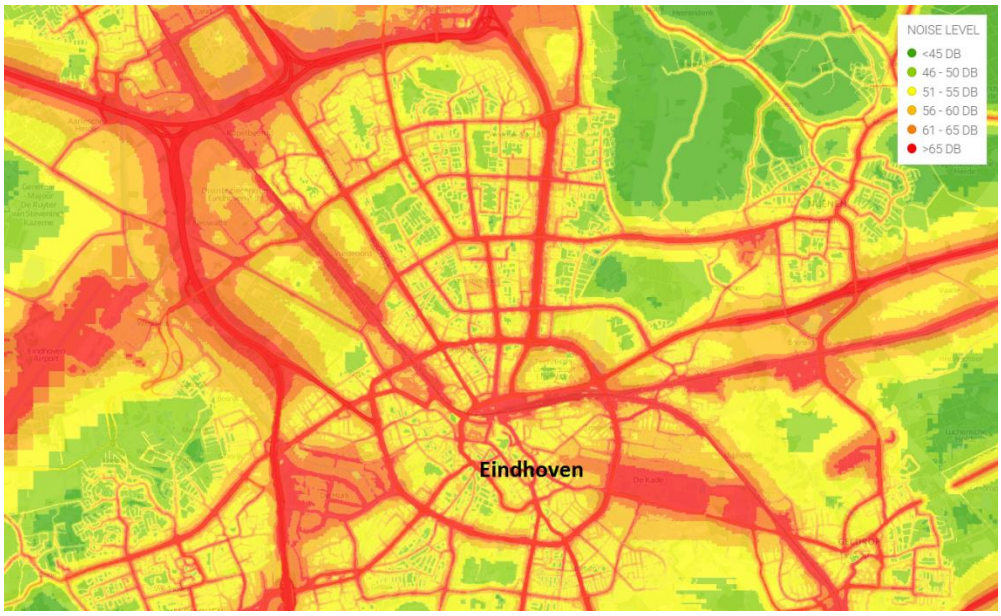


Figure 3.4: An example of noise pollution in and around Eindhoven (Spotzi, 2016)

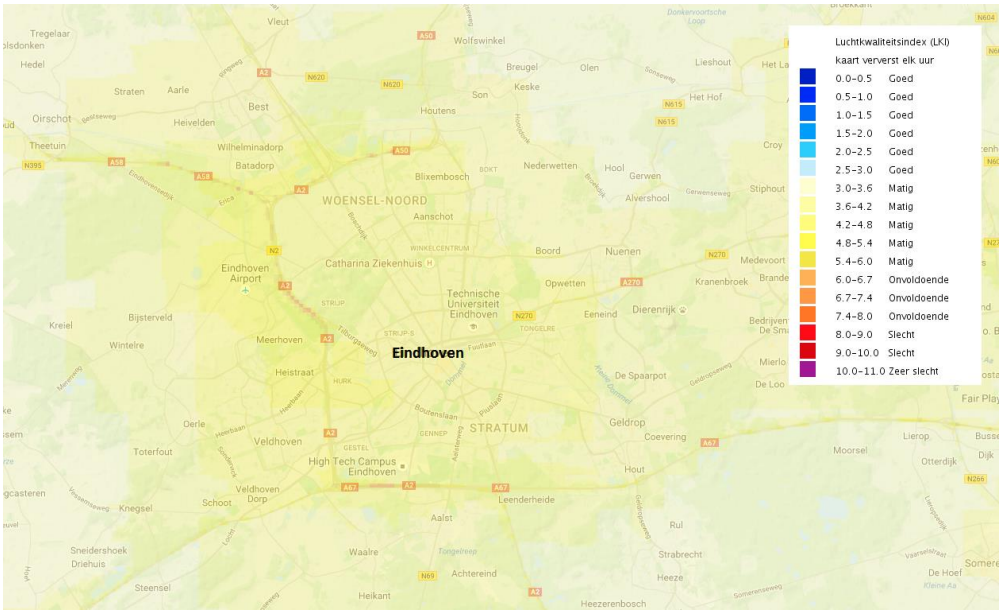


Figure 3.5: Air quality for the city of Eindhoven and surrounding areas (Luchtmeetnet, 2016)

3.3.3 Data Quality

When working with geographical datasets in a research there are always some problems that can arise concerning the data that need to be dealt with. First of all, data can be changed or corrupted without this being the intention. This can be easily mitigated by making back-ups of all data and changes. Secondly, it is always a possibility that the necessary data is either not available at all or is difficult to obtain. Each case of this problem arising is unique and therefore requires a specific way to deal with it that cannot be completely anticipated. One possible workaround is to use a similar dataset that is obtainable but might not cater to the needs of the research as much. The most inherent impediment of working with data, however, is data quality. It is an indication of how 'good' the data is and it is an ever present issue in GIS. Heywood, Cornelius and Carver (2011, p. 340) describe the unavoidability of data quality issues as:

"We must accept that no matter how careful we are in the preparation of our data and how cautious we are in our choice of analysis, errors will find their way into the GIS database"

They also define four data quality issues that could be present in any dataset used for this thesis:

1. *Errors* are flaws in the data that indicate a physical difference between the data and the real world. Errors can be singular but also can also be widespread throughout the dataset.
2. The *accuracy* of data indicates the differences between the estimated values and the true values. While datasets cannot be 100% accurate, they can be within a predetermined accuracy threshold. This links back to the MAUP and UGCoP.
3. *Precision* is the level of detail present in the dataset. A high level of precision does not mean the data also has a high level of accuracy.
4. If a consistent error throughout the dataset exists, it is called *bias*. An offset of 10 meters for every observation in a dataset is an example of bias (Heywood, Cornelius & Carver, 2011).

These problems with the data quality can have a multitude of sources. The can come from the source data, human intervention, incorrect selection of analysis operations, computational errors, incorrect transition between software or an incorrect presentation of results (Heywood, Cornelius & Carver, 2011). To test the data quality of the datasets used in this thesis, four parameters are identified that together should give a complete overview of it:

1. The data should be *complete*. This means that it should contain all the required information for the area it represents.
2. Considering that multiple datasets are used and they shift back and forth between different software applications, the *compatibility* of the datasets is important.
3. To ensure the compatibility of datasets it is important that they are *consistent*. Attributes need to be uniform, coordinate systems need to be uniform or can be transformed if not.
4. The final parameter is the *applicability* of the data. This parameter indicates if datasets are suitable to put through commands or analyses. It also indicates the appropriateness of the variables used in an analysis to solve the project's problems (Heywood, Cornelius & Carver, 2011).

The results of the data quality testing for all datasets used in this thesis can be found in appendix 2 at the end of the thesis report.

3.3.4 Data Preprocessing

A part of research methodology often needed when working with geographical data is data preprocessing. Only incidentally the data needed for research is gathered and constructed in such a way that a user does not need to process the data before doing analyses with it. The preprocessing of the data entails the process of manipulating data to a point where it can be used for analysis, in this case spatial influence modelling (see paragraph 3.2). While the datasets have some differences in the way they need to be preprocessed, only the general steps are discussed in this section.

A main benefit of preprocessing will often be that the data size is decreased. The most important preprocessing step for this is to clip all datasets to the research area. The result of this is that only the features remain that could be used in the analyses. For some of the more sizable datasets, such as the road networks used, this is taken a step further as all useless attribute information is also excluded from the datasets, same with all irrelevant features within the research are. Optimizing the size of the used data also leads to a decrease of the processing time of the analyses.

Besides these main preprocessing steps, some more specific, but necessary, preparations include the steps that need to be taken to avoid unnecessary errors during analysis. A good example is to make sure that only acceptable symbols are used in the datasets. Using letters with additions such as a diaeresis are a good example of how an analysis can suddenly and unexpectedly fail.

3.4 Enriching the Routes

Enriching the routes is the process of adding the spatial influence information to them. This is done in three distinctive parts; modeling the spatial influences with the use of geographical data and through one of the three discussed spatial influence modeling methods, adding this information to the GPS tracks and weighting the influence values to get valid results. This paragraph serves as an introduction to how these steps are approached.

3.4.1 Map-Matching

In relation to the MAUP and UGCoP discussed in paragraph 2.3, an important subject of the methodology that is still undiscussed is map-matching. The first related problem is how to implement it when integrating route data in the analysis. Part of the runners' dataset used in this thesis consists of GPS tracks of practice routes (see paragraph 3.3). To assess the influence of the spatial factors on runners, these tracks have to be matched to an existing street network. The question arises, however, how to do this to ensure reliability and accuracy of the matching of the tracks on the network. This process is called map-matching. Newson and Krumm (2009, p. 1) define map-matching as:

"Map-matching is the procedure for determining which road a vehicle is on using data from sensors."

Different Map-Matching Methods

In the paper Newson and Krumm (2009) give some options on how this might work. The simplest algorithm that could be used is to match each data point from a sensor to the nearest road. Given the inherent offset most sensors have, however, this method is deemed unreliable, especially if the road network is dense around the observations. To counter the errors that can occur when matching to the nearest road, modern algorithms look at sequences of points first before deciding on a match. A more advanced method proposed is the Hidden Markov Model [HMM]. A great advantage of the HMM over other algorithms is that it considers the connectivity of the roads and it considers multiple different paths at the same time before deciding on the most probable one. This leads to a more optimal balance between the route suggested by the location data and the feasibility of the path. The networks' connectivity is the most important attribute when working with the HMM. Before the algorithm decides on a transition from one road segment to another, it considers each possible connected road segment and assigns a probability to each based on their position opposed to the position of the next data point. Not only the next point is considered, however, as also the previous points are as they have an influence on which road segment is the most probable route. An observation made by Newsom and Krumm (2009) is that they found that the most probable route between two data points is often the one where the distance traveled over the route is the closest to what they call the great circle distance. This is another name for surface distance. This might seem obvious, but when previous and following points are taken into account, the real path could very well not be the shortest path, especially on complex street networks.

To speed up the computing speed of the HMM on the large road network used in their study, Newson and Krumm (2009) used the Viterbi algorithm. This algorithm is used to find the most probable sequence of Hidden Markov states or segments to form the most probable route (Forney Jr., 2005). The sequence

of the most likely HMM segments is called a Viterbi Path. This dynamic programming algorithm is often used when trying to find the right road segment for location points, for example GPS measurements (Cova, et al., 2008).

A classic example of a method used for map-matching is the Dijkstra algorithm. The original principle of the algorithm is to find the shortest path from point A to point B. The idea initially stems from 1959 and since then multiple modifications to it have been proposed. For this algorithm to be relevant to the thesis, these modifications are necessary as:

"However, this algorithm is not efficient for searching shortest path in large graphs" (Rodríguez-Puente & Lazo-Cortés, 2013, pp. 1-2)

While the original algorithm could be used on simple square networks, it would probably not suffice on more complex networks, such as a contemporary road network. A lot of the modifications that have been added throughout the years focus on the hierarchy within the road network (Rodríguez-Puente & Lazo-Cortés, 2013). Examples of this are the method proposed by Gonzalez et al. (2007), which assumes that users want to travel on the largest roads, or the method proposed by Geisberger et al. (2008), which only uses the road segments (edges) that are related to 'important' intersections (nodes). The question is, however, to what extent these motorized vehicle aimed extensions are relevant to the needs and wants of runners. A second branch of modifications to the algorithm focusses on reducing the runtime of the algorithm or the computational costs. An example of this is the proposed modification to the A* modification of the Dijkstra algorithm by Rodríguez-Puente and Lazo-Cortés (2013). This modification tweaks the spatial complexity of the Dijkstra algorithm instead of the temporal complexity, which the A* modification does. A major drawback of these algorithms, however, is that they do not consider the previous measurements in the chain when creating the shortest path, something the HMM does. Therefore the Dijkstra algorithm and all its modifications keep their focus on the shortest path instead of having the focus on the most logical path. As a result, the HMM and Viterbi algorithm are chosen over a version of the Dijkstra algorithm.

By map-matching the GPS routes, adding the spatial influence factors on the network to them is possible, leading to the spatial influences happening on a more accurate location than if the measurement errors of GPS are not dealt with (Newson & Krumm, 2009).

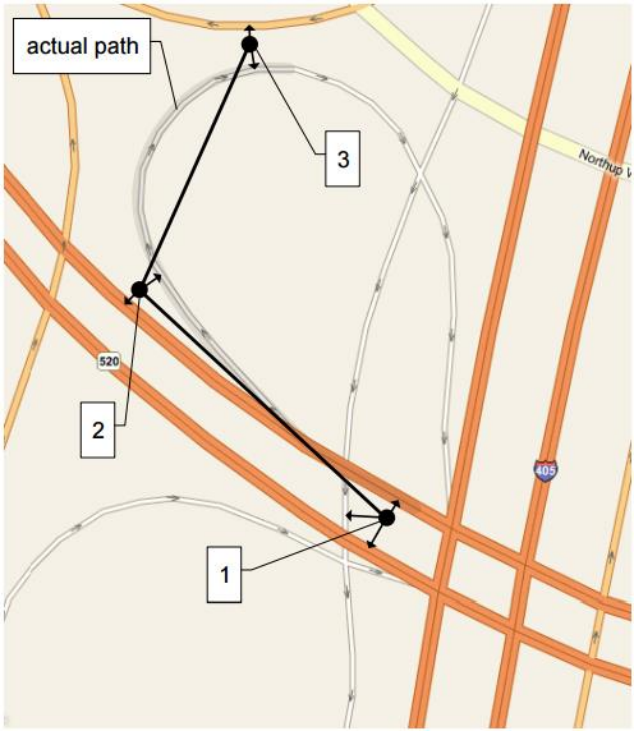


Figure 3.6: An example of choosing the logical route with a HMM (Newson & Krumm, 2009, p. 1)

How is Map-Matching done in this Research?

As mentioned, the HMM as proposed by Newson and Krumm (2009) is used. With this measure the most likely route is decided on when going through a set of GPS positions. As said in the previous section, the HMM method considers not only the possible road network options, but also the previous points on the GPS track to determine the most logical route. A good example of this can be seen in figure 3.6. While GPS measurement 3 seems to be closer to the orange road above it, it still snaps to the highway exit below it. When an algorithm only considers the distance to a road segment to determine the probability of it being used, point 3 would have been snapped to the wrong segment. By also considering the route before it, the logical conclusion is that the highway exit is the correct road segment for the measurement. The Viterbi algorithm is used in conjunction to the HMM to calculate the optimal path with the optimal product for the measurement and transition probabilities of the GPS tracks.

If this method is used without stating any restrictions, it would calculate a probability for each road segment in the used network file for each GPS measurement. This is an unnecessary consumption of time and processing power when doing the calculations. To mitigate this, a distance restriction is added to the HMM. When a road segment is more than 50 meter away from the GPS measurement, it is not taken into account as a possible road segment for a runner to be on. Newson and Krumm (2009) used a restriction of 200 meter for their case study concerning a driving car. It is argued here, however, that considering that GPS measurements should have a below 7,8 meter accuracy in 95% of the measurements and the higher density of the network available to runners over cars, this threshold of 200 meter is deemed undue for runners. However, they did experience problems with some measurements that could not snap to a road segment within 200 meter in their case study. Reason for this seemed to be that the car would be in a parking garage not on the network or when a tunnel or

urban canyon resulted in extreme GPS noise (Newson & Krumm, 2009). Both these explanations are not expected to be encountered when studying runners. An explanation for when a map-matching error does happen in this research could be that the runner is running off-road.

Besides the distance restriction, three other possible restrictions could be set into place. The first is to exclude low probability routes. What is meant by this is that when the route distance differentiates too much from the great circle distance (straight surface line), the probability of the route is set to zero. Considering runners do not necessarily take the fastest possible route from their start to their end point and that they can have the same start as end point (Allen-Collinson, 2008; Ettema, 2015), this restriction does not seem useful for this thesis. The second possible restriction is to exclude GPS outliers. These outliers happen when the route requires the vehicle to travel at speeds greatly exceeding the allowed speed limit or above a certain maximum speed value (Newson & Krumm, 2009). This restriction is focused on motorized vehicles, however, and not on human runners and therefore not needed in this thesis.

The fourth restriction proposed by Newson and Krumm (2009) provides a bit of a challenge for this thesis. They argue that when GPS measurements are too close together, it is uncertain if the movement comes from actual movement or from GPS noise. To mitigate this problem they determined that two GPS measurements should be at least twice the standard deviation of the GPS measurements from each other to validate the movement being the result of actual movement. This standard deviation of GPS measurements is the average error in GPS accuracy of all measurements. The problem it poses for this thesis is that while it is probable that the GPS sensors used by the runners in the data are about as accurate as the sensors in the case study of Newson and Krumm (2009), the average speed at which the route is travelled should be lower for a runner than for a car. As a result, it is expected that measurements being within this constraint are more frequent among runners than they are among motorized vehicles. Therefore, it is decided to only implement this fourth constriction if there is reason to believe the integrity of the research results is affected because of not implementing it.

A good example of the visual results of map-matching can be seen in figure 3.7. A track can be seen, which is the connection of all GPS measurements (red track). This track, however, does not logically flow when taking the road network into account, as it goes from one side of the road to the other. As explained, this is due to the inherent GPS noise (Newson & Krumm, 2009). After map-matching the route, however, it follows the road normally, as visualized by the black line in figure 3.7.

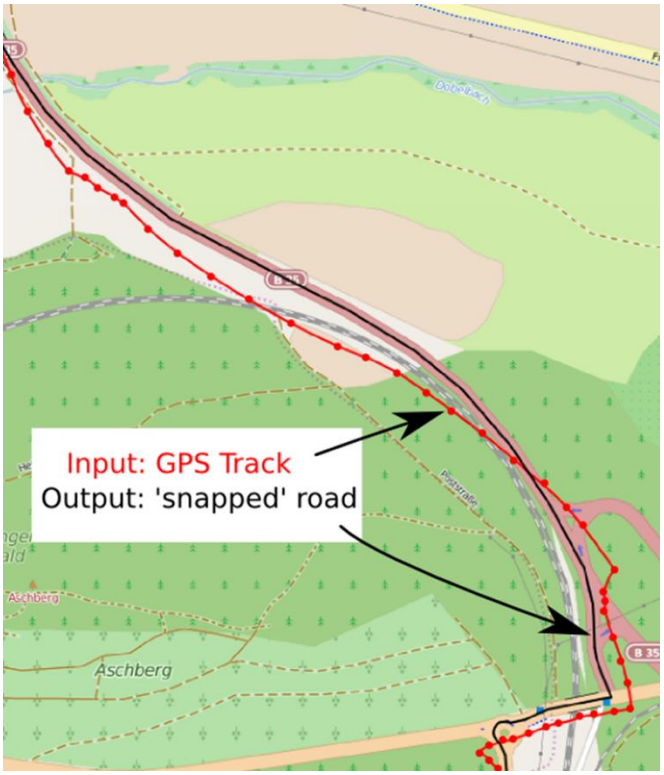


Figure 3.7: Example of map-matching a GPS track (openstreetmap.org, 2015)

Computational Approach

The script that is used to perform the map-matching through the Viterbi algorithm is called '2mapmatcher.py'. The draft of the script was built by Simon Scheider (2016) and has been edited since then to suit the GPS and network data used in this research. The script is divided into three chunks of Python definitions; preprocessing the GPS tracks and extracting a file list from them, the map-matching itself, exporting the Viterbi paths. These chunks are briefly explained below.

The first part of the script is used to select 200 GPS tracks from the route file and split them into separate files. This is done by generating 200 random numbers between 0 and 21405 (the amount of GPS tracks in the file) and storing them into a list. The list is then used to extract the corresponding tracks from the input route GeoJSON and store them in a new GeoJSON file. These 200 routes are used throughout the whole analysis, as using a sample reduces the processing power necessary to complete the methodology. With an input file containing the 200 tracks, the following step is to split them into 200 separate datasets. This is done by using a unique attribute per track (the ID of the run) to construct unique names for the datasets and storing 1 track with a matching ID per dataset. Additionally, the coordinates are transformed from the global spatial reference system WGS1984 to the Dutch spatial reference system RD New and the data type is changed to an esri Shapefile. This latter part and the map-matching itself are executed per 10 tracks, as it requires a lot of internal memory to execute it.

The second part of the script entails the map-matching of the 200 tracks to the network dataset used in this thesis. This process requires five input parameters, explained in table 3.2 below. The workflow of map-matching one track works as such:

1. The two endpoints and the length of each segment in the road network are extracted and put into arrays.
2. With the network information, a network graph is constructed of the largest connected component in the network. This does exclude some network segments on the edges of the research area that are cut off from a connection to the graph because of the network being clipped by the research area.
3. The x and y coordinates of all point measurements in a track are appended to an empty list.
4. The segment candidates for the first point in the track are gathered using the network and the Euclidean distance the point is from the segments.
5. The segment candidates for every point after the first one are gathered, while taking into account the candidates for the previous points.
6. The Viterbi path is determined by gathering a sequence from all segments with the highest probabilities.
7. The segments with the highest probability have their unique ID extracted to the GPS measurement it corresponds to.
8. The most probable path is cleaned up by removing redundant segments, which happens because multiple points have same segment as most probable candidate.

Parameter	Explanation
Track	The esri Shapefile containing the 200 GPS tracks.
Segments	The esri Shapefile containing the network that is used. This is a combination of the two network datasets introduced in paragraph 3.3 and include only the segments a runner is allowed to run on.
Decay constant of the network	The network distance after which the match probability of a segment to a point falls below 0,34, which signifies how far GPS point are situated from one another. This is calculated with an exponential decay function.
Decay constant of the Euclidean distance	The Euclidean distance after which the match probability of a segment to a point falls below 0,34, which signifies how far a GPS point can deviate from its true position. This is calculated with an exponential decay function.
Max distance	The maximum Euclidean distance a candidate segment for map-matching can be from the GPS point.

Table 3.2: Input parameters for map-matching

Finally, the path of most probable segments is exported to a new shapefile. The end product consists of 200 shapefiles that each contain a multi polyline feature that is the route the runner took on the network. As said before, a reason why this process might fail along the way is when a runner went off-road. This could result in a lack of segment candidates within the maximum distance from the point, which returns an error.

3.4.2 Adding the Influences to the GPS Tracks

The results of the spatial influence modeling can be used to enrich the GPS routes once the map-matching of them is done. With these the effectiveness of the chosen methods can be determined. As the results of the three spatial influence methods discussed in paragraph 3.2 are stored differently, they too require a separate approach for how to combine them with the GPS measurements.

The spatial influence factors that are modeled with the surface based component are stored in the street network the GPS points are matched to. This means that to couple an influence value with a certain GPS point, an attribute value of the street network needs to be called upon and stored in the GPS point attribute table.

For each influence factor using the distance based component, a spatial influence function is determined to model the influence probability decay over distance. With this function, the influence probability and value can be mapped to the GPS point attribute table.

When using an area based component to model an influence factor the process is similar to the previous one. If distance is used to model the influence probability of all influence sources, a spatial influence function to model probability decay needs to be determined. If the probability is based on for example the size of the source, the size of the sources inside the influence area need to be calculated and added to the source attribute table. Thereafter, the necessary influence probabilities can be appended to the attribute table of the GPS measurements.

Computational Approach Surface Based Influence

To utilize the surface based influences in the Python tool, a standardized computational approach is needed for the relevant influence factors. As mentioned in section 3.3.1, the surface based component is the most straightforward one, which in turn translates to the most straightforward computational approach.

1. Determine the variables in the dataset that are needed to calculate the influence value.
2. Influence values are assigned to the different unique values of the variables and stored in a new variable. When necessary, multiple variables are combined to come to an influence value.
3. Determine if the GPS point is on an influence source. When this is the road network, it is done by map-matching the GPS routes to the network (see previous paragraph) and then extracting the influence value assigned to the road segment to the GPS point. When it concerns polygons, the value of the influence source is directly copied to the GPS point on the source.
4. This influence value is stored in a new attribute per GPS measurement on the route to be further manipulated in another part of the analysis.

Computational Approach Distance Based Influence

To utilize the distance based approaches, a standardized computational approach is needed for the exponential influence decay and linear influence decay. The variable part of the approach comes from how strong or lasting the used influence factor is estimated in this thesis when taking the distance between measurement and influence source into account.

The different influence sources of the input influence factor are split from each other and stored in separate files using the Fiona and OGR Python packages. Key in this process is that the definition that is created to do this operates in a flexible and automated way. A Python definition is written, called 'valueSplitter'. To achieve the flexibility and automation, the input variables for it are used to open the input file, select the correct influence variable and build and populate new data files with distinct and logical names and file paths. These input variables are always stated in the definition wherein the 'valueSplitter' is called. The basic structure of it is:

1. Call 'valueSplitter' with stated input variables.
2. The input file is opened.
3. The unique influence values are extracted into a list and sorted from low to high (-1 to 1).
4. New data files are written for the present unique influence value and populated with the first feature of the original file to get the correct data schema.
5. The filenames are appended into a list and serve as the return value of the definition.
6. The features of the input file are then stored in the data file that corresponds with the influence value of the feature.
7. The first feature that was used to get the correct schema in the output files is deleted from the files, as it is either a duplicate of another feature or an incorrect feature for the file.

When all the features are stored into the correct datasets, these datasets are opened with the return values from the 'valueSplitter' definition and a Euclidean distance is calculated within the research area. Thereafter the Euclidean distance is reclassified with an exponential or linear decay function, such as seen in equation 3.1 and 3.2. The strongest influence probability (closest to a source), decay rate of it and the 0% probability distance is dependent on the input features.

To get the influence on the track an influence raster is made that represents the reclassified Euclidean distance, which is the influence probability. Thereafter, the cell values are extracted to an empty attribute of the GPS measurements that are on them and multiplied by the influence value of the source.

Computational Approach Area Based Influence

The computational approach for the area based influences is to some extent similar to the approach for the distance based influences. Up until the point where the Euclidean distance is reclassified for the different influence sources in the same influence factor to either exponential or linear decay, the general computational approach stays the same. The way the influences are thereafter manipulated is different, however.

1. Influence rasters are constructed for all different kind of influence sources that fall under one influence factor by reclassifying the influence probability based on the distance between source and GPS measurement.
2. The value of the raster cell (influence probability) the GPS measurement is on is extracted to the point for each influence raster and multiplied by the influence value of the source to get an influence per kind of source on the runner.
3. The influences of the different rasters are added up to get a cumulative influence value for the influence factor.

3.4.2 Influence Aggregation

A problem that is encountered when generating influence values for the routes and each GPS measurement on the routes is that the results can be biased towards certain measurements. This happens when the distances between GPS measurements are not equal. These distances can be measured in space or time. When this is the case, the input value from each GPS measurement towards the value for the entire route needs to be weighted. This done by ways of the following function:

(5)

$$\frac{\sum_{i=0}^{n} W_j * V_j}{\sum_{i=0}^{n} W_t}$$

W_j *= Weight of measurement j based on the Δd of the measurement*
Δd = The distance between a measurement and the previous one
V_j *= The influence value of measurement j*
W_t *= The cumulative weight of all measurements in a route*

The weight value W_j is based on the Δd of a measurement. This is the distance between the measurement and its predecessor. By using this input as weights, measurements with a large distance between it and the previous one will weight heavier in the cumulative influence value W_t. In contrast, when measurements are relatively close together in distance, they will have a decreased weight in the final cumulative influence value.

A problem that needs to be addressed is how to tackle the firs GPS measurement of the route, as it does not have a Δd. The solution used in this thesis is to assign an average weight to this measurement. When a route consists of 50 GPS measurements, the Δd of 49 of them will result in a total weight of 1. The first measurement will get a weight of 1/49, resulting in a W_t of around 1,02.

Computation Approach

The Pythonic approach to weight the influence values of the different influence factors can be found in the script '2WeightedAverages.py' in the companion data. It consists of two definitions of which the first is used to assign the weight of the first value and calculate the linear distance between all the points. The linear distance instead of the network distance is used for this as it simplifies the process and is the clear method in which the sum of the Δd is the same as the total distance. As GPS measurements are relatively close together with runners, the linear distance is expected to not differ too much from the network distance and in general, be a bit lower. There is a check in place to assess if the network distance and the linear distance differ too much. The second definition is used to assign the weights of every point after the first one. The approach of the script is as follows:

1. A track is opened in read only mode.
2. Attributes are added to the data scheme to store weight and Δd in.

3. The amount of point measurements in a track are calculated by looping through the points.
4. A writer file is constructed with the expanded data scheme and the points are looped through again.
5. For the first point the average weight is assigned and the covered distance is set to zero.
6. For every other point, the Δd is calculated by getting the distance between its coordinates and the coordinates of the previous point.
7. The Δd is stored in an attribute for each point and the covered distance is calculated by adding up every point distance to it.
8. This information is written to the writer file and the total covered linear distance and the point count of the route are used as return values.
9. A check is performed to see if the linear distance falls within the safe margin of 10% of the actual track distance, which is an existing attribute in the track.
10. The weight of all points after the first one are calculated by dividing the Δd by the total linear distance and the expected W_t of 1 plus the average weight assigned to the first point.
11. The weight of all points after the first one are also written to the writer file.

3.5 Validation of Results

The final phase of the research methodology is the evaluation and validation of the results gathered in the previous parts of the analyses. Result evaluation and validation is an important part of any project or research (Saltelli & Annoni, 2010). First, different methods of research validation and evaluation are discussed, after which the chosen method in this research is explained in detail.

3.5.1 Methods for Validation

There are multiple ways in which data validation and evaluation can be performed when working with geographical data. Throughout the academic literature, there is a strong emphasis in testing the sensitivity of the output or results when working with models to this end. By doing a sensitivity analysis the origin of uncertainty found in the model output can be pinpointed. When dealing with multiple error sources, which is quite probable, doing a sensitivity analysis helps in getting a clear picture of how each error influences the results. Based on the outcome of the analysis, it might be necessary to consider adjusting the model to filter out the sources of some of the errors (Saltelli & Annoni, 2010).

Another method often integrated in GIScience is to combine online data with offline data. This entails that on one hand GIS is used to analyze geographical data. The results of the analysis is thereafter evaluated by looking at qualitative data to assess the parallels and differences. This method often applies to research in which subject behavior plays a part (González-Bailón, 2013). An example of how this could apply to this thesis subject is if the runners used to gather the GPS tracks also gave interviews on why they chose to run certain routes or distances.

As a slightly different approach to evaluate if the research results seem to conform to reality is to statistically test them. Statistics can be used to determine the extent to which the results can explain

the behavior of runners in general. In GIScience research, statistical methods are often used to test the fitness of the research results (Eman, Brown & AbdelMalik, 2009; Austin, e.a., 2016). With a dataset containing 21405 GPS tracks, it is possible to compare the research results to the density of runner tracks in an area. This way the behavior of the runners can be analyzed without the need for qualitative data, as discussed before. If the correct influence factors are chosen and modeled, it should mean that areas with a high density of runner tracks also have a more runner friendly environment than areas with a low density of runner tracks. As qualitative data on the runners from the survey is not available to the researcher, a statistical analysis is used to evaluate the research results.

3.5.2 Regression Analysis

The statistical type of analysis used in this thesis is the regression analysis. Regression is used to estimate the relationship among variables. This relationship is between a dependent variable (density of runner tracks) and the independent variables (influence factors). The use of a regression analysis is twofold. Firstly, it predicts the effect the independent variables, or predictors, each have on the dependent variable and secondly, it can be used to assess how much of the variance in the dependent variables is explained by the predictors (Xin & Xiougang, 2009). As mentioned in the previous section, regression is useful to determine to what extent the selection of influence factors and influence modeling methods used in this thesis are fit to explain where runners are active.

While many different regression models have been formulated since its inception in the early 19th century by Legendre (1805) and Gauss (1809), it is rudimental to discuss all of them in this paragraph. The models relevant for this thesis conform to the linear regression model category. This model assumes a linear relationship between the dependent variable and its predictors. When one predictor is used, the analysis is called a simple linear regression, but when multiple predictors are integrated, as with this thesis, the multiple linear regression is used (Xin & Xiougang, 2009). Cook and Weisberg (1982) decades ago already defined the base formula used for the multiple regression as:

(6) $$Y = b_0 + b_1X_1 + b_2X_2 \ldots + e$$

Y *= dependent variable*
b_0 *= the constant where the function meets the Y-axis*
$b_N X_N$ *= the influence of influence factor N on the dependent variable*
e *= the unobservable error*

When using the linear regression methods, there are a couple of assumptions the model has to conform to. When these are not met, it is hard to argue the regression results as an approximation of reality (Denuit, Mesfioui & Trufin, 2019). The assumptions are:

1. The dependent and independent variables need to be scalar. While it is possible to integrate a categorical variable as independent variable, this is usually not done.
2. The dependent variable is influenced by all the independent variables.

3. The relationship between de dependent variable and its predictors is linear. Firstly, this is assumed and tested after completing the regression analysis by comparing the predicted values for the dependent variable with its residues.
4. The residues need to have a homoscedastic distribution. This means that the noise in the relationship between the dependent variable and its predictors is the same across all values of the predictors.
5. The residues follow a normal distribution. This means that for all the combinations of values of the predictors, the values of the dependent variable have a normal distribution.
6. There must not be any autocorrelation between the independent variables. If there is autocorrelation, it means two or more independent variables try to explain the same thing (Mardikyan & Darcan, 2006).

These six assumptions and the execution of the regression analysis are discussed in great detail in chapter 7 of this thesis.

3.6 Workflow

The methodological workflow consists of six phases aimed at reaching the research objectives and answering the research questions. Below, the six working phases are discussed briefly. Important to note is that the workflow is not completely parallel to the order in which the remainder of this thesis is structured. The thesis structure has already been discussed in paragraph 1.7.

1. Literature Exploration: To gather information on which influences need to be modeled when researching runners, the scientific literature is explored. The results of this, the influence factors, can be found in paragraph 2.2. Besides the influence factors, literature on problems when modeling spatial influences and the operationalization approaches is gathered. These subjects have already been discussed in paragraph 2.3 and this chapter.
2. Data Exploration: This phase comprehends the assessment of useful data and the acquisition of it. The different influence factors require different datasets to model. Paragraph 3.3 discussed which datasets are used to model each factor.
3. Spatial influence modeling methods: In paragraph 3.2, a distinction between three different spatial influence modeling methods was discussed. The surface based, distance based and areal based each require different computational approaches to model an influence factor. Which approach is best suited for which factors is discussed in chapter 5 and is dependent on both the influence factor and the used data.
4. Route enrichment: Once all spatial influences are modelled, the GPS routes are enriched with the influences from each influence factor. Per factor and modeling method this is done in a different way. In Paragraph 3.4, the general method of the route enrichment is explained. Chapter 6 comprehends a more detailed explanation of the route enrichment and the discussion of the preliminary results.
5. Influence aggregation: The next step in the process is to aggregate the influences per GPS measurement. To counter any differences in the distances between measurements, the influences are aggregated to represent the influences on the part of the route between the

measurement and the previous one. The methods behind it are discussed in paragraph 3.4 and the results are discussed in chapter 6.

6. Regression analysis: To test the fitness of the model of factors and spatial influence modeling methods, a regression analysis is performed. The results show how much of the variance in where runners run is explained by the chosen factors and modeling methods. The regression analysis and its results is discussed in greater detail in chapter 7.

Chapter 4
Data Exploration

4.1 Data Structure

In this chapter, the primary data sources, consisting of the GPS tracks of runners, is explored in a descriptive manner. Firstly, a quick introduction of the data format and its content is given in this paragraph, after which the following paragraphs focus on descriptive statistics for the entire dataset and the 200 GPS tracks used for the analysis (see paragraph 3.4).

As mentioned in the previous chapter, the GPS tracks are stored in a GeoJSON file, which the geographical version of a JSON file (JavaScript Object Notation). It is used to store both the spatial and non-spatial attributes of geographic features. The complete file is categorized as a feature collection, with each GPS track representing a feature within it. The spatial information of the GPS track is stored within the geometry of the feature as a LineString feature based on a list of x and y coordinates for the GPS measurements of the route. The feature connects the GPS measurements through the shortest distance between the points and does not follow any road network yet.

The non-spatial information of the tracks are stored in the properties of the feature. This information includes a wide range of quantitative information about the track, but nothing about the runner. Table 4.1 includes all the attributes from the track and a short explanation of each.

Attribute name	Explanation	Example
Geometry		
Type	The geometry type of the feature.	LineString
Coordinates	A list with x, y coordinate pairs.	[(x, y), (x,y), (x,y)]
Properties		
TOD		
averageSpeed	The average speed of the run in m/s.	4,324
Day	Which day of the week it is.	0 = Monday, 6 = Sunday
Daylight	If the run was with daylight.	0 = no, 1 = yes
Distance	The distance of the run in meters.	7598,56
Duration	The duration of the run in seconds.	1800 (30 minutes)
effectiveTime	The duration of movement on the run in seconds.	1770
Hour	The hour of the day the run started.	8 = between 8 AM and 9 AM

Month	In which month the run took place in.	0 = January, 11 = December
Runid	A unique identifier for the run.	vZg9JoCnlP
Season	Which season the run took place in.	0 = Winter, 3 = Fall
startTime		
Training	Another identifier, which is missing in some cases.	8gRlfurZ7K
Week		
Year	In which year the run took place.	2015

Table 4.1: Structure of the GeoJSON file

4.2 The Population

As mentioned before, the non-spatial information that is included in the GeoJSON is only about the run itself, and not the runner. To get an image of how the population is constructed, this paragraph explores some descriptive statistics of the population, which is all 21405 routes in the dataset. Of the attributes discussed in the previous paragraph, not all are relevant to this data exploration. The duration, average speed and distance attributes are explored first, after which some temporal attributes, such as the month, the day of the week and the time of day are discussed.

Table 4.2 shows the descriptive statistics of the aforementioned attributes duration, distance and average speed. The average duration of all the runs is 1909 seconds, which is a little over half an hour. With all the runs, the runners ran on average 4342 meter, with even exceptions running more than 21 kilometers. While this research does nog differentiate between casual runners and more advanced runners, these statistics do show that both groups are represented in this dataset. The runs were on average performed with 2,35 meter per second, which is around 8,48 kilometer per hour. With the maximum run speed of 18 kilometers per hour, some fast runs were also executed. When looking at the standard deviation [σ] and the coefficient of variation [CV] of these attributes, it is clear to see the runners have a much greater variation in how long and far they run, than how fast they run during the route.

Attribute	Count	Min	Max	Mean	σ	CV
Duration (sec)	21405	318	7178	1909,458	745,138	0,390
Distance (m)	21405	1000,080	21522,209	4342,160	1927,190	0,444
Average speed (m/s)	21405	1,381	4,994	2,355	0,412	0,175

Table 4.2: Descriptive statistics of the population

As mentioned, exploring the distributions of runs in a temporal manner might also provide some interesting insights. Figure 4.1 shows distribution histograms of the runs in the months, days in the week and time of the day. When looking at the monthly division, the amount of runs stays relatively equal, with the exception of considerably less runs being performed in February and a small peak in September. The daily division does give some interesting insight as Monday is by far the most popular day of the week to run on. Expectations would be to see an increase in the amount of runs on Saturdays and Sundays, which even have the lowest amount of runs being performed during them.

The hourly division shows a clear pattern of runs being performed either in the morning of evening after dinner. No runs in the population dataset were performed between 12AM and 6AM, with the peak of runs being ran between 10PM and 11PM.

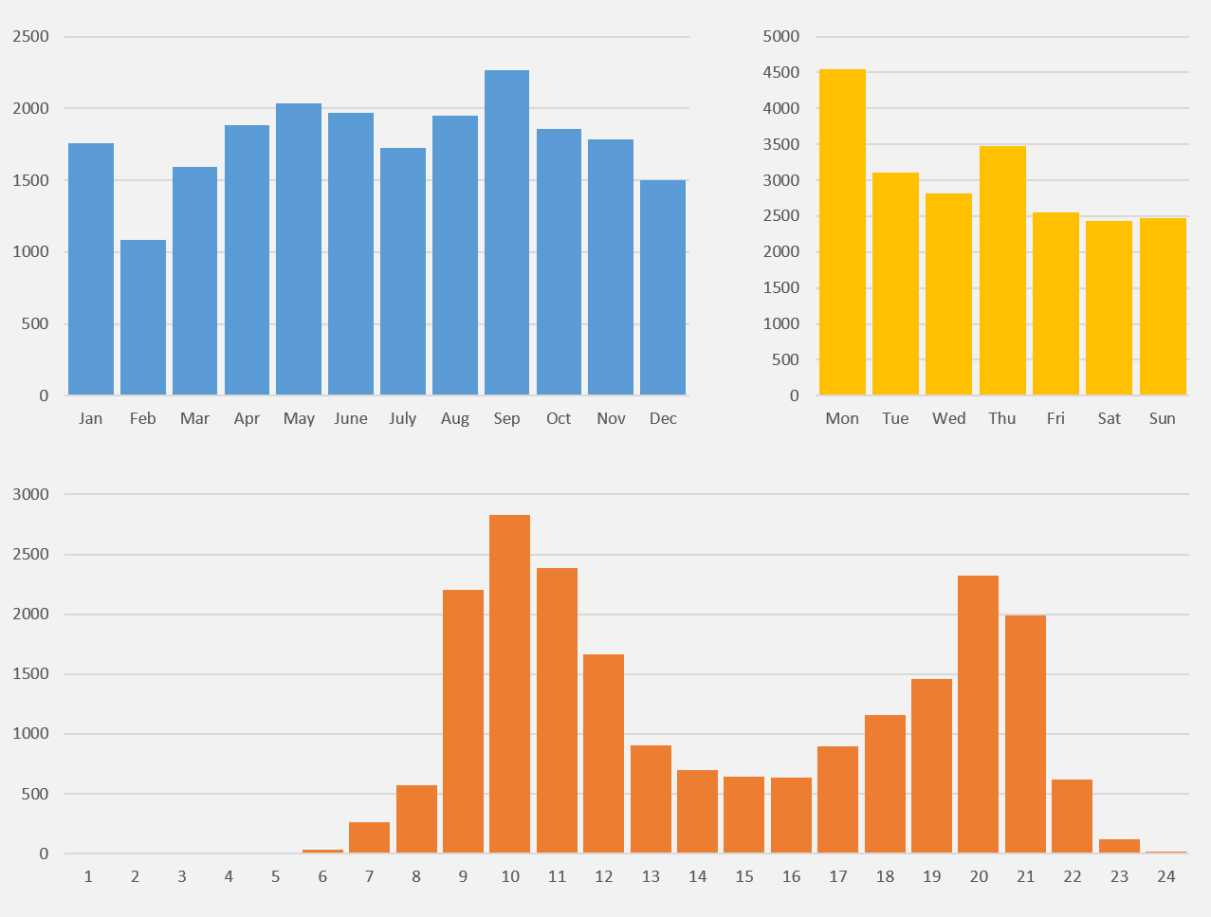


Figure 4.1: Temporal division of runs over the months (blue), days of the week (yellow) and the hours of the day (red)

4.3 The Research Subjects

When looking at the duration, distance and average speeds of the runs that are taken as sample in comparison to the statistics of the population, it is expected that the variation (not the CV) for these attributes is less than when looking at the population. This is due to the fact that when you only take around 1% of the entire population at random, it can be expected that the extreme outliers are not part of the sample tracks. When looking at table 8.3, however, it is noticeable that the min and max values for these attributes are not too different from the same values of the population. Exceptions are the minimum duration being more than twice as long and the maximum distance of a route being around 5 kilometers shorter. This observation is reflected in the minimal differences in the CV between population and sample, together indication that for speed, distance and duration, the sample seems a good representation of the population.

Attribute	Count	Min	Max	Mean	σ	CV
Duration (sec)	200	736	6575	2049,495	902,251	0,440
Distance (m)	200	1460,162	16031,581	4589,716	2266,374	0,494
Average speed (m/s)	200	1,388	4,374	2,316	0,381	0,165

Table 4.3: Descriptive statistics of the research routes

When looking at the temporal divisions of the 200 research subject runs on a monthly, weekly and hourly basis, considering only a random 1% of the population runs is used, it is expected that patterns seen in the temporal divisions of the population are more randomized. While this seems to be the case for the monthly and weekly division, the hourly division still shows the same patterns as the hourly division of the population. Of the research subjects, most were completed either in the morning or in the evening after dinner, with the peak being between 10AM and 11AM.

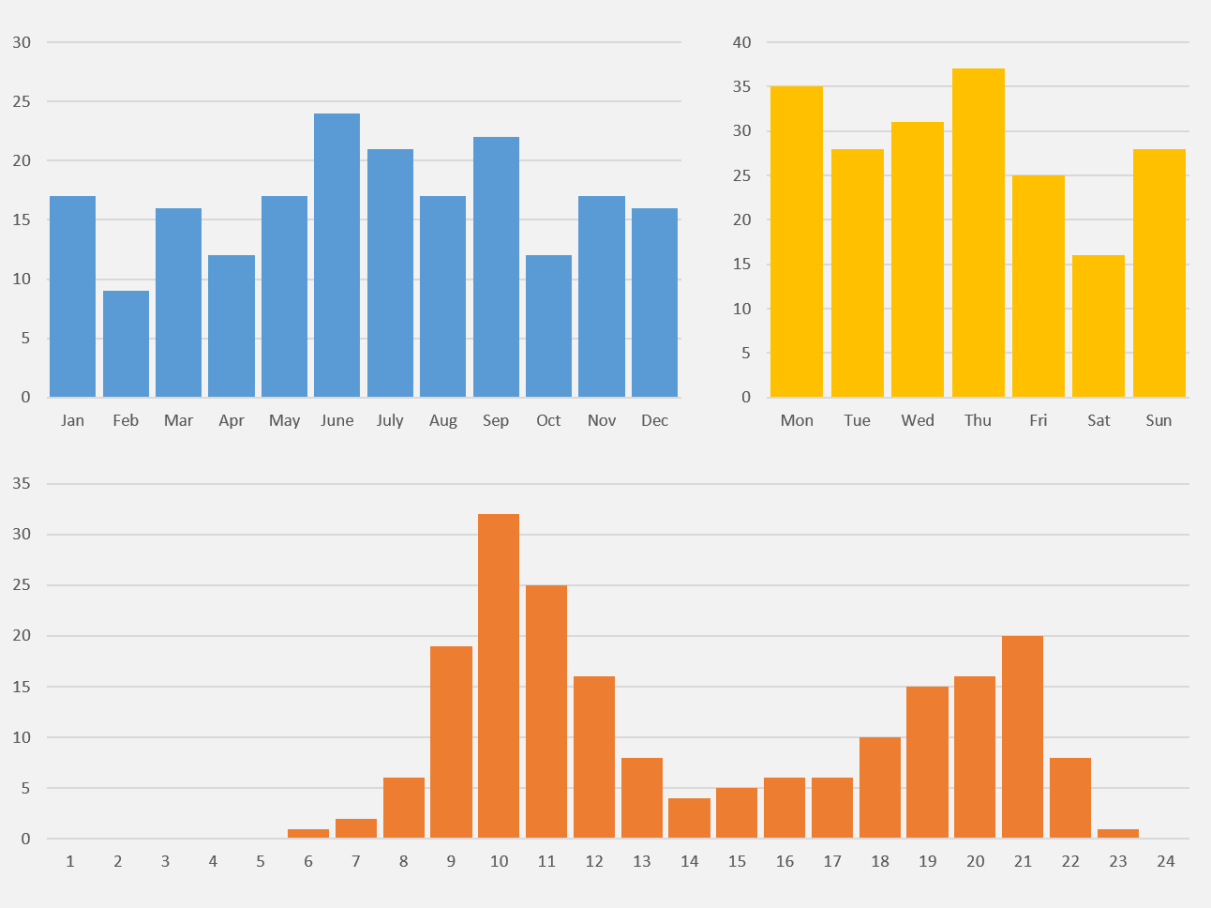


Figure 4.2: Temporal division of runs over the months (blue), days of the week (yellow) and the hours of the day (red)

Chapter 5
Spatial Influence Modeling

5.1 Introduction

In this chapter the preliminary results from the route enrichment are discussed per influence factor. In addition, it is discussed which spatial influence modeling method (see paragraph 3.2) is used for which factor, and why. This decision is made based on both what the factor represents, but is also determined by the data that is used to model it. The Python script(s) used to model the influence factor can be found in the companion data to this thesis and are mentioned at the end of each influence factor in this chapter.

Surface based influences are assigned quite easily, as they require that a runner needs to be on the influence source to experience it. This makes the probability of an influence binary. The distance and areal based methods provide more of a challenge to correctly assign. Table 5.1 shows the different criteria a factor and its data can have to ideally be modeled by either the distance based method or the area based methods.

Generally the distance based influence modeling is used when the factor can be modeled with data that contains a singular kind of influence source per dataset used and few influence sources, while the areal based influence modeling is used for factors for which the data contains either multiple kinds of influence sources or has a lot of sources in general. Besides that, distance based methods use the closest influence source to get the influence probability, while the probability of multiple sources or the maximum probability can be used when using the area based methods.

Criteria	Distance based influence modeling	Area based influence modeling
Different kind of influence sources	One	More than one
N amount of influence sources	Moderate	A lot
Data type	Polyline, polygon	Point, polygon
Influence probability	Closest source	Multiple sources or max value

Table 5.1: Differences distance and area based methods

5.2 Surface Based Factors

5.2.1 Running Surface

The first influence factor modeled with surface based influence is the running surface. As the data used in this thesis for this influence factor is a road network, and the factor representing the ground the runner is on, it seems only logical that this factor is modeled as a surface based influence. The network dataset contains multiple attributes that provide information on the quality of the surface. A mentioned in paragraph 2.2, the hardness and smoothness of the road, together with the presence of hindrances on the road together comprehend this factor according to previous work. The relevant attributes are listed below in table 5.2, together with the possible values and how they are reclassified for this research.

Attribute	Explanantion	Possible values
verhardingstype	The road surface type of the network link.	- verhard (-1,0) - onbekend + gemend verkeer(-1,0) - onbekend + fietsers, bromfietsers (0,0) - onbekend + voetgangers (1,0) - halfverhard (0,5) - onverhard (1,0)
hinder	The degree of hindrance on the network based on size of the road and main occupant	- <2 meter + gemengd verkeer (-1,0) - <2 meter + fietsers, bromfietsers (-0,3) - <2 meter + voetgangers (0,4) - 2-4 meter + gemengd verkeer (-0,8) - 2-4 meter + fietsers, bromfietsers (-0,1) - 2-4 meter + voetgangers (0,6) - 4-7 meter + gemengd verkeer (-0,6) - 4-7 meter + fietsers, bromfietsers (0,1) - 4-7 meter + voetgangers (0,8) - >7 meter + gemengd verkeer (-0,4) - >7 meter + fietsers, bromfietsers (0,3) - >7 meter + voetgangers (1)

Table 5.2: Surface quality influence value mappings

There is a problem that needs to be addressed concerning the usage of these datasets. The difference in precise and relevant attributes and values in the network dataset. While, as can be seen in table 5.2, there is an attribute to model the hardness of the roads in this dataset. Around 15,5% of the roads in the network dataset have an unknown value for hardness of the road. In these cases, the main users of the road segment are included to assign the influence value, as roads with an unknown hardness and pedestrians as users are generally unpaved and roads with an unknown hardness and mixed occupants are generally paved. The smoothness and hindrances are more difficult to gather from the data, however. The hindrance is approximated using both how wide a road is and who its main occupants are, information that is in the dataset. The general rule with this that a combination of runner friendly users and a wider path provide the least hindrance to a runner. This is seen as separate to the social safety factors, as this comprehends the expected amount of room available to a runner.

This problem that comes from data limitations needs to be taken into account when evaluating the results of this influence factor in the next chapter. Before a route is enriched with surface quality information, the reclassified attribute values are summarized and standardized with the following formula for the used network:

(7) $$0{,}5 * surface\ type + 0{,}5 * hindrance$$

Python scripts: 0surfaceQualityPreprocessing_fast.py, 0mapmatcher_fast.py, 1surfaceQualityInfluence.py

5.2.2 Verbal Harassment

Verbal harassment is modeled using criminality rates from the CBS over the course of 2 years, 2016 and 2017. These numbers are available on a neighborhood scale, which means that analysis cannot be done on a smaller scale to avoid the MAUP. However, ideally this influence factor would modeled through the distance or area based influence methods. Because the scale cannot be smaller than the neighborhood scale, this is sadly not possible. This leaves the verbal harassment to be modelled through the surface based method, which means that the influence a runner experiences is based on in which neighborhood he or she is running.

Since the criminality rates from 2016 and 2017 are used, a spatial files containing the neighborhoods from the same year is used to avoid the problem of reclassified neighborhood divisions in the research area. The criminality rates that are openly available through the CBS are subdivided into three different categories of crime:

1. Sexual offenses and harassment.
2. Violence and disturbance of public order.
3. Stealing and looting (CBS, 2019).

The numbers only include registered crimes, which reasonably means the actual number of crimes is still higher. Of the three categories mentioned above, the first two are relevant for modeling the verbal harassment, as they represent negative influences from one or more persons on another one in the public space, in which the runners move. The third category is discarded and not used in this analysis.

The file containing the criminality rates is merged with the neighborhood file based on a certain standardized neighborhood code in the files. With this method, all 270 neighborhoods in the research area are correctly merged and now contain rates on sexual offences, harassment, violence and disturbance of public order. These criminality rates are now just the absolute amount of registered crimes though. To study the effect on runners, however, they have to be standardized to correct for the skewed division of crime rates between neighborhoods. This can be done in multiple ways. The population of the areal division is often used to standardize crime rates, so you get X amount of crimes per N inhabitants. For runners, using the population makes less sense, as this influence factor should tell the influence of verbal harassment through crime rates, and should not include the influence of the population of neighborhoods. Therefore, the area of a Neighborhood is used. This way the crime rates represent X amount of crimes per square kilometer.

The last step to model this influence factor is to reclassify the standardized crime rates to a 0 to 10 scale, in which 10 represents the most negative influence for this factor, and 0 represents no influence. The rate of 100 crimes per square kilometer per year, meaning 200 over the research data of 2016 and 2017, is chosen as the point for assigning a 10 to a neighborhood. Figure 5.1 shows the results of this reclassification. A quantile classification is chosen for the map to as optimally show where the areas with the least and most negative influence are situated. As can be expected, the urban areas, especially of the largest city, Eindhoven, have a stronger negative influence, while the rural neighborhoods seem to score a low negative influence.

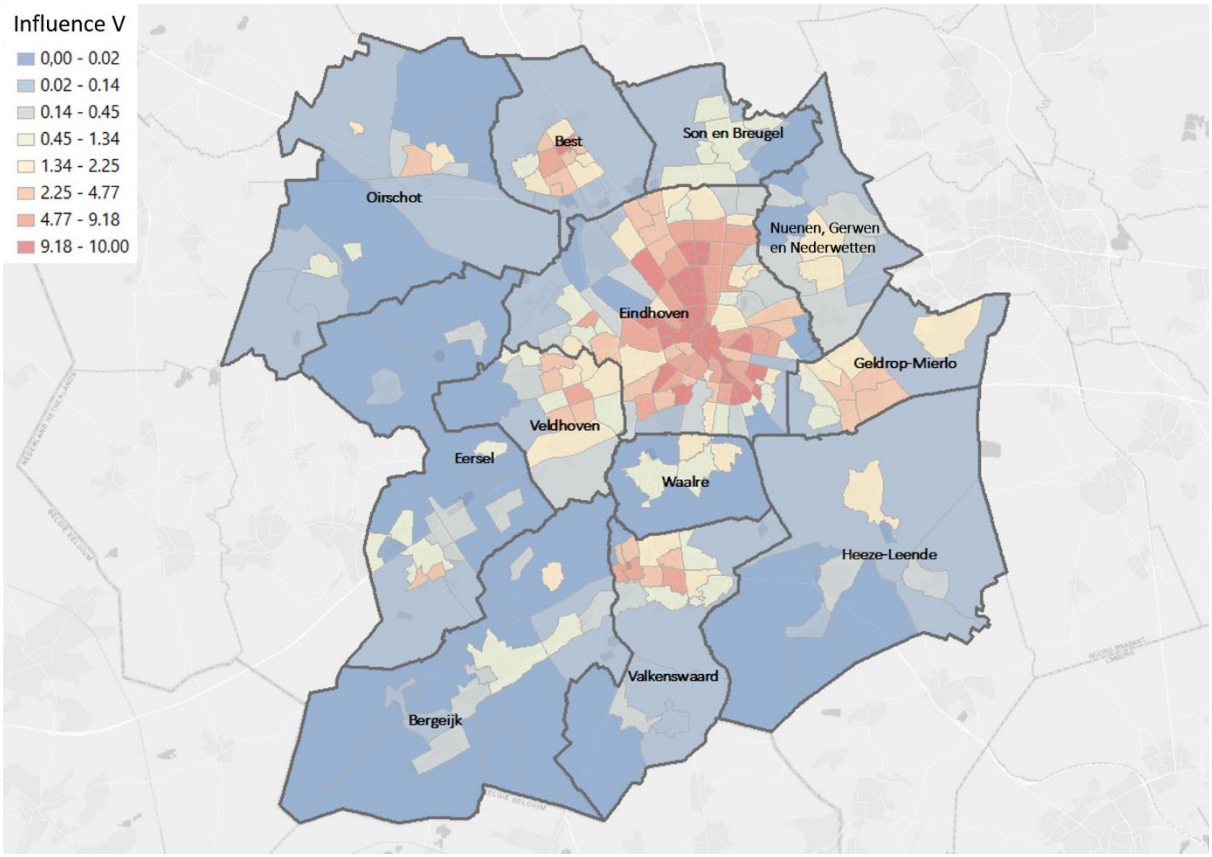


Figure 5.1: Visualization of verbal harassment influence in the research area, 2016-2017

Python scripts: 1verbalHarassmentInfluence.py

5.2.3 Street Lighting

To model the influence street lighting has on a runner, a decision had to be made about what data to use to tackle this factor. The municipality of Eindhoven has a lot of open source geographical data, including a dataset with all the street lighting within the municipality (Gemeente Eindhoven, 2019). While this dataset would be ideal to use, the surrounding municipalities lack such a dataset. Therefore, the dataset with the bicycle only network is used (see section 3.3.2), as it contains an attribute on the amount of street lighting on a network link. As this network does not include the paths exclusively used

by pedestrians (and runners), the assumption is made that these paths do not contain street lighting, except when close to a network link that does contain it.

Table 5.3 shows the values the street lighting attribute in the network dataset can have and which influence values are given to them. To assess if a runner is within the area of effect of street lighting, buffers are built around the road network. If the GPS measurement is within that buffer, the influence value within the road network is appended to the GPS measurement attribute table of the route. The reason this influence factor is not modeled through the map-matching of the routes, as is the case with the surface quality, is because a source of lighting does not have to necessarily be on the road the road is running on.

Lighting value	Influence value	Buffer diameter (meters)
Good	1,0	20
Average	1,0	10
Bad	-1,0	20
Unknown	0,0	1

Table 5.3: Street lighting influence value mapping

A difference between the surface modeling between this influence factor and the other two influence factors, is that if there is 'no influence probability', meaning the runner is too far away from light sources, the influence probability is actually still 1, but the influence value has dropped from 1 to -1. It is seen as unsafe to run on roads that are not properly lid when there is no daylight (see Addy, et al., 2004; Lee & Moudon, 2008). However, when a run is performed with daylight, this influence factor is negated by assigning influence values of 0. This is for street lighting the no influence probability scenario, which is temporal instead of spatial.

Python scripts: 1streetLightingInfluence.py

5.3 Distance Based Factors

5.3.1 Motorized Vehicles

The Dutch government has openly available data on the traffic load on the road network of the past decennia. The corresponding data size, however, makes it impossible to analyze this data for the researcher when the information of more than a couple of days is used. To find an alternative method, different road datasets are compared on the relevant information they held. The two best candidates for analyzing the presence of motorized vehicles and its effect on the social safety of runners are either the amount of traffic lanes or the maximum speed of the road. As the amount of traffic lanes did not provide consistent and truthful information available for the entire network, the maximum speed is used for this influence factor. The maximum speeds in the research area are spread between 15 and 120 kilometers per hour. Figure 5.2 shows the division of the maximum speeds in the research area.

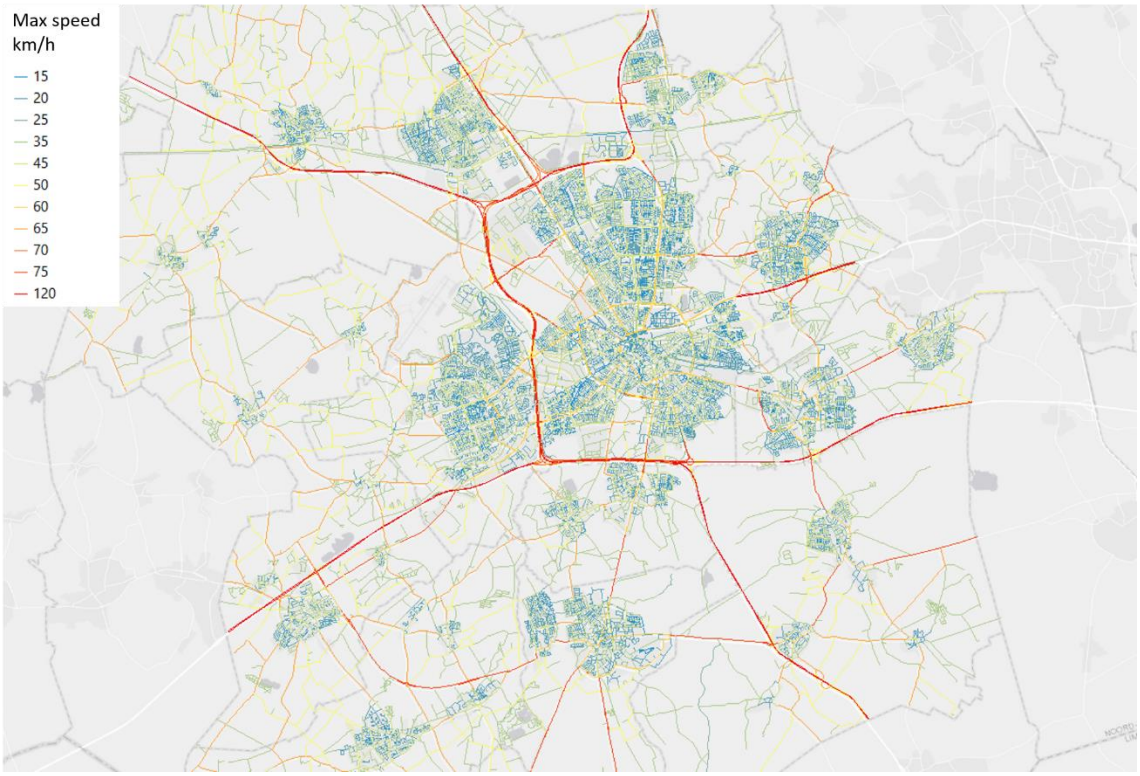


Figure 5.2: Overview of the maximum speeds on the road network in the research area

To model the influence of the motorized vehicles, the distance between each GPS measurement of the route and the nearest road segment is calculated and reclassified into an influence probability. This is done through a linear decay function, of which the steepness is based on the maximum speed of the closest road segment. The decay function for all different maximum speeds is:

(8) $$P_x = 100 - \frac{100}{speed/10} * Distance_x$$

The influence value remains consistent at -1,0, with the probability of the influence happening being the variable value, as discussed in paragraph 3.2. The reach of the influence, through its probability of happening, is bigger when the maximum speed is higher. Using the aforementioned equation, a road with a maximum speed of 80 kilometer per hour has a P of 0 when the runner is 8 meters or further away.

Python scripts: 1motorizedVehiclesInfluence.py

5.3.2 Cyclists

As is the case with modeling the motorized vehicles, the data with information on the traffic load of the road network is too big to process, meaning an alternative needs to be found. The road network that is

also used for the map-matching and modeling of the surface quality also includes indications if cyclists are allowed on the road segments. The influence the cyclists have on a runner can be gathered from the distance the runner is from the nearest road segment. For this influence factor, there are two other considerations to take into account. The first is that the decay rate is the same for all the roads, as cyclists are not separated on the basis of how fast they can go, such as was the case for motorized vehicles. The decay function utilizes linear decay, with the threshold for an influence probability of 0 set to 5 meter. The decay function can be seen below in equation 9a. While in reality, this threshold should be lower, the inherent errors the GPS measurements contain require a less strict threshold. The second consideration is that, to assess the amount of cyclists, another measurement needs to be taken into account. This measurement determines the influence value each road segment where cyclists are allowed gets. The population density on a neighborhood level is used for this. Neighborhoods with a higher population density are also expected to attract more cyclists. In the second function (9b), the calculation of the influence value through the population density is shown.

(9a) $$P_x = 1.0 - 0.2 * Distance_x$$

(9b) $$V_x = 0.0001 * Density * P_x$$

The population density is gathered from the Statistics Netherlands neighborhood dataset that is also used with the verbal harassment influence factor. The density is extracted by checking if the point geometry of the GPS measurement overlaps with one of the neighborhood polygon geometries and storing the population density of the neighborhood in the attribute table of the point.

Python scripts: 1bicycleInluence.py

5.4 Area Based Factors

5.4.1 Natural Areas

As mentioned in paragraph 5.1, the area based spatial influence modeling is used for influence factors with different kinds of influence sources and many influence sources. The amount of different sources close to a runner at any given time, make it impossible to assume the runner is only influenced by one of them. In the case of the natural areas, a subdivision is made between green nature and blue nature, or water. Thereafter the green nature and water is subdivided into the different types that the Top10NL dataset provides. Table 5.4 shows these subdivisions and which influence values they are given for the analysis.

The influence values assigned to the green nature are done based on variety, height of the upper tree line openness and average size of this natural polygon. This means that both mixed forests and heaths get a high influence score, while lines of poplar trees get a lower score. Cemeteries are seen as off limit green nature and therefore receive a negative score. Different polygons of water are assigned scores based on their use. If the purpose of the water body is industrial, such as docks or a water treatment plant, it is assigned a negative score, while if its purpose is recreational or unknown, it receives a positive

score. With the polyline features for water, the influence score is assigned based on how wide the waterway is.

The natural areas spatial influences are modeled through exponential decay, rather than linear decay, due to the probability that there are obstructions between the runner and the source when the distance increases. The three formulas below show the decay functions for respectively green nature, water bodies and waterways. The difference is the value for the decay rate (see section 3.2.2). Waterways have the highest value for the decay rate (-0,15), which means it is the steepest. These differences are meant to simulate the difference in influence sources that rise up from ground level (green nature) and sources that lie lower than the runner (water).

Dataset	Feature type	Possible values	Influence value
Green nature			
Top10_NL_TERREIN_VLAK	Polygon	- Gemengd bos	1,0
		- Loofbos	1,0
		- Heide	1,0
		- Naaldbos	0,8
		- Zand	0,6
		- Populieren	0,6
		- Bos: griend	0,2
		- Dodenakker met bos	-1.0
Water			
Top10NL_WATERDEEL_VLAK	Polygon	- Overig	1,0
		- Natuurbad	1,0
		- Vistrap	1,0
		- Drinkwaterbekken	0,5
		- Haven	-1,0
		- Viskwekerij	-1,0
		- Vloeiveld	-1,0
		- Waterzuivering	-1,0
Top10NL_WATERDEEL_LIJN	Polyline	- 3 – 6 meter	0,6
		- 0,5 – 3 meter	0,3

Table 5.4: Influence value mapping of natural areas and water

(10a) $$V_{dist=x} = 100 * e^{-0,05x} * V_{dist=0}$$

(10b) $$V_{dist=x} = 100 * e^{-0,10x} * V_{dist=0}$$

(10c) $$V_{dist=x} = 100 * e^{-0,15x} * V_{dist=0}$$

The actual modeling part of this influence factors, means splitting up the different influence values for a source from table 5.4 and storing them in different shapefiles. Thereafter the Euclidean distance is calculated within the research area and reclassified using the aforementioned influence functions. The result is a 5 by 5 meter raster with the reclassified values, which represent the influence value on the distance of the cell value from the influence sources.

The final step to finalize the influence raster by combining the separate rasters into one. Important for this is that the cell sizes and positions are the same for all the influence rasters. Using the research area as the boundaries for all of them and utilizing a consistent 5 x 5 meter raster cell size avoids misalignment problems.

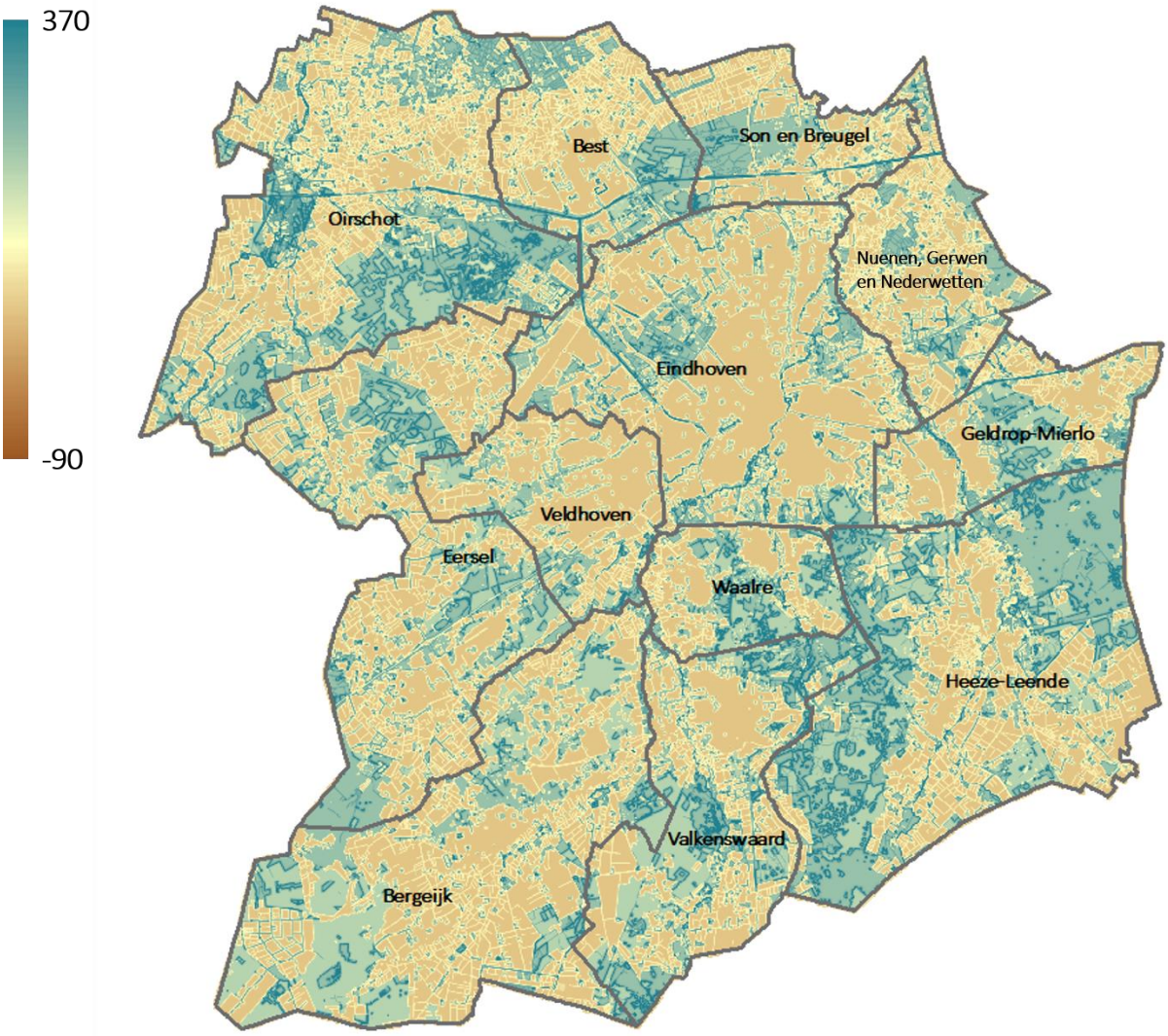


Figure 5.3: Distribution of influence values in the research area

Figure 5.3 shows the combined influence raster, consisting of ten separate influence rasters. With the calculation of the influence raster values between -90 and 370 can be found in the research area. This means that when standardized, the influence values for this factor will range between -0,9 and 1,0, as every value above 1 is decreased to 1. The majority of the low values in the research area are due to either it being an urban area or farmland, two land uses not included in this analysis that each take up a decent amount of land.

Python scripts: 1naturalInfluence.py

5.4.2 Sound Pollution

The data to model the sound pollution is gathered from the RIVM (2017) and contains and combines sound pollution numbers from highways, other roads, railways, airports and airplanes, industry and wind turbines. This information is gathered in the period between 2011 and 2016 and while being gathered at different scale, it is now stored in an asci grid of 10 by 10 meters. An assessment of a possible MAUP is therefore necessary. As discussed in paragraph 2.3, Menon (2012) argued that the used areal units must be the same in shape, size and neighboring structure. While the shape and neighboring structure might be the same in an asci grid, the cell sizes of the combined datasets do not seem to be the same, as can be seen in figure 5.4. Area A at Eindhoven Airport clearly uses data (airports and airplanes) with bigger cell sizes than the 10 by 10 grid. Area B at the A2 highway that runs in between Eindhoven Airport and the city of Eindhoven does seem to use sound pollution data (highways and other roads) that conforms to the 10 by 10 grid.

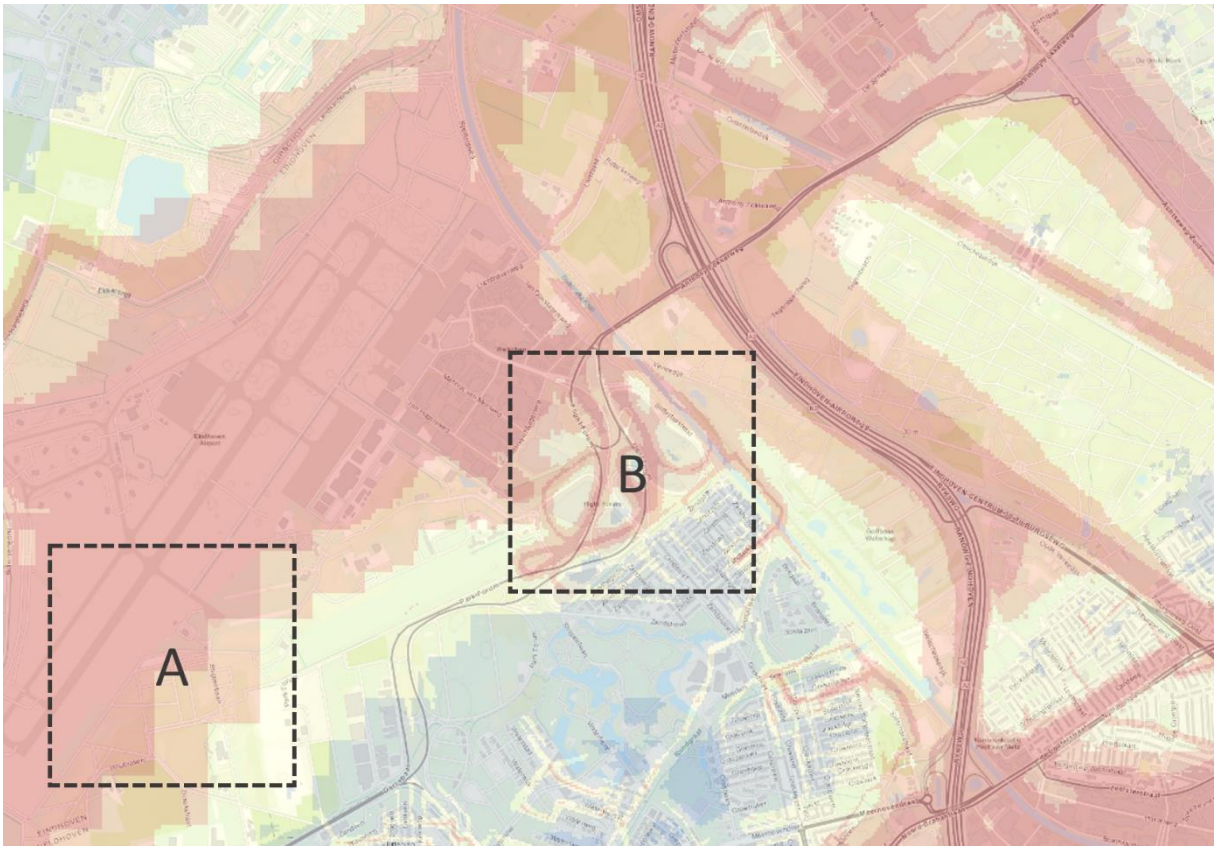


Figure 5.4: Different sound pollution cell sizes around Eindhoven Airports

Table 5.5 shows the classification the RIVM made for this dataset in the meta information. This classification is also utilized for this thesis, as their calculation is based on the desirable and allowed standards in the Netherlands (Atlas Leefomgeving, 2018b).

Value RIVM	Sound level	Classification RIVM	Influence value
1	< 45 dB	Very good	1,0
2	45 – 50 dB	Good	0,6
3	50 – 55 dB	Reasonable	0,2
4	55 – 60 dB	Moderate	-0,2
5	60 – 65 dB	Bad	-0,6
6, 7, 8	> 65 dB	Very bad	-1,0

Table 5.5: Influence value mapping for sound pollution

An issue regarding this influence factor the researcher ran in to, was that the asci grid dataset was too big to process with Python, as it could not be loaded. To work around the large data size, ArcMap from the ArcGIS package was used to convert the dataset into a raster file and FME is used to clip the raster file to the research area beforehand to reduce its size. Thereafter the cell values are reclassified to a range from -1,0 to 1,0 conform the values in table 5.5. Considerably more raster cells have positive values than negative. With this information, the average influence value for a raster cell in the research area is 0,38. While expectations could be that runners in general have a greater chance to run in a cell with a positive influence value for this factor than negative, it is hypothesized the opposite is true. Runners often move across an existing road network, and while they should use paths suited for them, these are expected to be in close proximity to sources of sound pollution such as roads accessible by motorized vehicles. A more in depth look into the route enrichment results can be found in chapter 6.

Influence value	Amount of cells abs.	Amount of cells %
-1,0	462.487	6,42
-0,6	509.650	7,07
-0,2	843.906	11,71
0,2	1.363.680	18,93
0,6	1.643.506	22,81
1,0	2.380.803	33,06
Total	7.204.032	100,00

Table 5.6: Descriptive statistics of sound pollution cell size values

Python scripts: 1soundPollutionInfluence.py

5.4.3 Air Pollution

As discussed in paragraph 3.3, four measures of air pollution are used to model this influence factor, bigger and smaller particles (PM10 and PM2.5), nitrogen dioxide (NO2) and soot (EC). These datasets are downloadable in a 25 by 25 meter raster format together with their metadata. An important note to make concerning the cell size of the raster is that these datasets are a combination of large scale air quality datasets with a 1 by 1 kilometer cell size, and locally gathered data on a 25 by 25 meter grid. As the air quality is measured in only a select number of locations throughout the country, the values in between measurement points are estimations (Atlas Leefomgeving, 2018a). This is a data limitation accepted by the researcher.

As the air quality is already estimated when taking sources and distances from sources in account, that part of the areal spatial influence modeling does not need to be performed for this influence factor, same as was the case with the previously discussed sound pollution. Instead, the focus in the modeling is determining influence value. The influence probability is always 1, as air is always something a runner is breathing. Table 5.7 shows the accepted yearly average concentration per pollution. These are measured in micrograms per cubic meter (µg/m3). These maximum average accepted concentration is taken as the -1 value for the individual pollution.

Pollution type	Allowed average concentration (µg/m3)
EC	25
PM2.5	25
PM10	40
NO2	40

Table 5.7: Average yearly accepted concentration per pollution in µg/m3 (Luchtmeetnet, 2019)

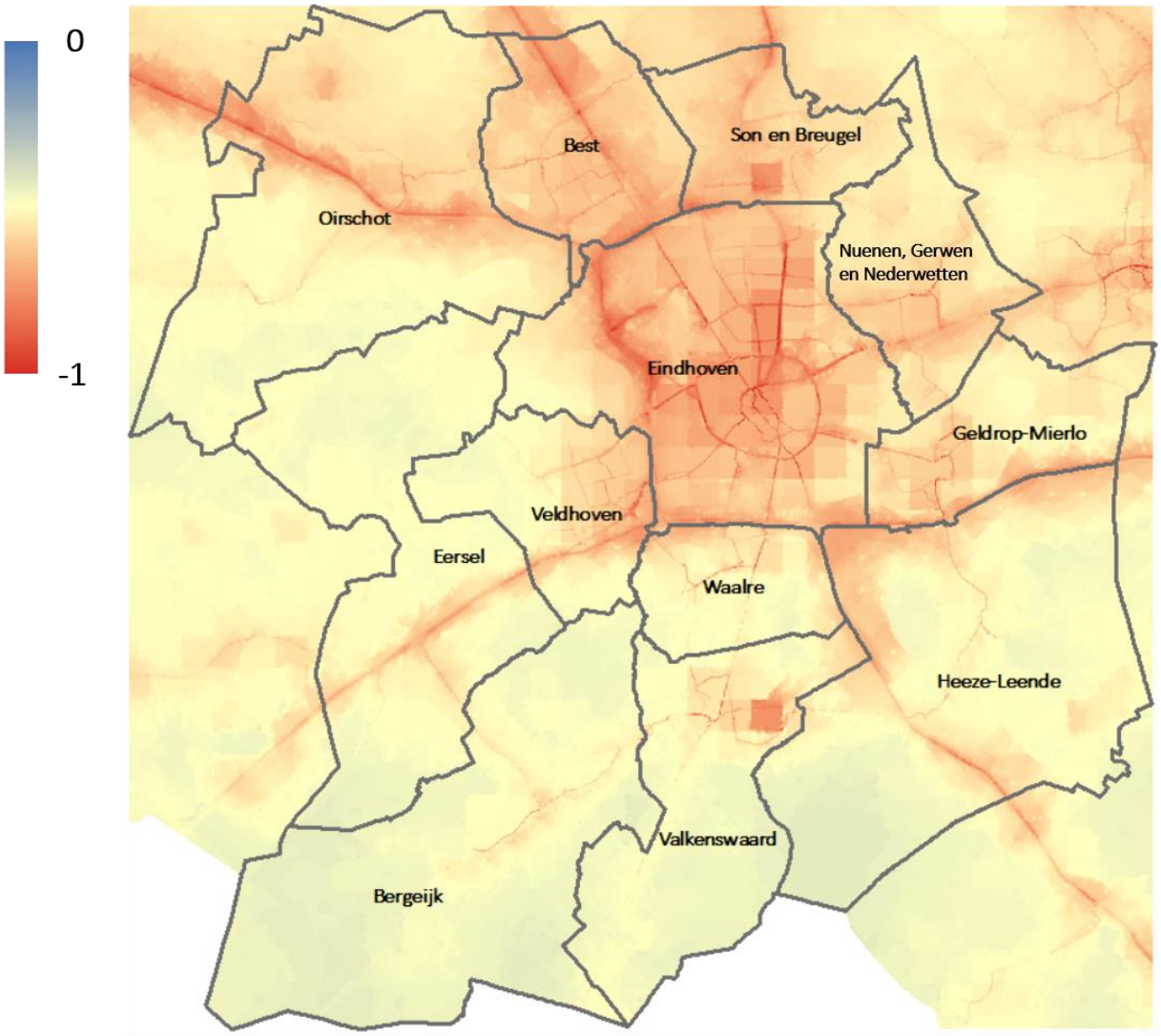


Figure 5.5: Air pollution in the research area

When combining the reclassified rasters, it is chosen to set the -1 value of the combined pollution to 50% of the values from table 5.7. The resulting influence raster of air pollution can be seen in figure 5.5, where an area without air pollution would have a blue color. As the figure shows, cells with values that low do not exist in the research area. The least negative influence of the air pollution on a runner in the research area is -0,46, while the most negative influence of a cell is -1. The large scale data used for the initial pollution rasters can be clearly seen in the figure, with the 1 by 1 kilometer cells providing strict borders between different influence values.

Python scripts: 1airPollutionInfluence.py

5.4.4 Variety in Surroundings

The final influence factor to be discussed is the variety in surroundings. As discussed in paragraph 2.2, runners prefer to run in a varied environment over a monotone environment. To assess the environmental variety for the runners, land use data is used. If a runner traverses an area with a lot of different land uses, this influence factor has a positive influence on the runner. To model this factor, a different approach is utilized than in the other factors in this paragraph. Wherein those started the modeling process in the datasets containing information on the influence factor, this factor is modeled with the GPS measurements as the start.

For each of the GPS measurements, a temporary buffer is created with a radius of 50 meter. Point by point, the polygon geometries of the land uses are counted if they are (partially) within the buffer of the point, after which the counted value is extracted to the GPS measurement. The result of this process is a clear image of which measurements are within a varied environment. Important to note, however, is that both 'positive' environments (nature, etc.) as 'negative' environments (large roads, etc.) are included in this analysis, as it is not an assessment of the variation of stimulating environments, but of all environment types.

Python scripts: 1environmentalVarietyInfluence.py

Chapter 6
Route Enrichment

6.1 Introduction

With the spatial influence factors being modeled, the next step is to enrich the runner tracks with the influences. Dependent on the factor, this can be done in multiple ways. The next two paragraphs go into detail on which methods are used for this and the benefits and problems of them. The map-matching of the runner tracks is given special attention in this chapter. Thereafter the influence aggregation is discussed briefly. Important is to discuss the decisions made for the aggregation. The last paragraph of this chapter includes some descriptive results of the route enrichment.

6.2 Map-Matching

As said, the map-matching gets some additional attention in this chapter, due to its complexity. The map-matching itself is used as a way to get the correct road segment of the used network dataset matched to the GPS measurements. As mentioned in paragraph 3.4, a method called the Hidden Markov Model and the Viterbi algorithm to connect the right road segments are used. To search for the correct road segment for each point, both the Euclidean distance from a point to the nearest segments and a network distance from one point on the road network to another are used. The result of this is a Hidden Markov State with probabilities of a GPS measurement being on a certain road segment. For this all road segments within a predetermined distance from the GPS measurement are used. With this scenario, the closest segment would always get the highest probability.

To make the map-matching process more logical, the Viterbi algorithm is used to fetch a path of most probable Hidden Markov States. These are not a sum of the closest segment for the GPS measurements, but rather the segments with the highest probabilities when also taking the next and previous GPS measurements into account. The resulting path should be the route of the runner along the road network.

6.2.1 Map-Matching Results

When the map-matching was performed on the 200 routes researched in this thesis, 5 routes failed to get a most probable path. The reason for this was that these routes were partially outside of the research area, which resulted in no segments being found within the maximum search distance for those routes, as the network data is clipped to the research area. The other 195 routes did return a most probable path.

The quality of the map-matched path does differ from route to route, however. Figure 6.1 shows the path and corresponding GPS measurements of the run with ID ‘pHdj1K7zEM’. This is an example of a route where the map-matching algorithm worked perfectly.

Figure 6.2: Example of successful map-matching for run with ID 'pHdj1K7zEM'

There are, however, some problems that impede on perfect map-matching of the routes, as a lot of them did not result in a complete route. Figure 6.2 shows the track of the run with ID ‘bQKBCvS6GM’. This run illustrates two of the problems. The first problem is the incompleteness of the road network, combined with the possibility of runners going off-road. While the road network is as complete as it gets, the road network in the Netherlands is ever changing, which results in road segments missing and the network being outdated still. On top of that, it only includes registered segments. It does not take into account that runners do not strictly follow the existing road network and can opt to cross from one segment to the other without there being a crossing. Area A in figure 6.2 shows this problem. The runner crossed across a short stretch of grass to get from one road segment to the other. The map-matching

algorithm utilizes a probability decay for road segment candidates that excludes the creation of Hidden Markov States if the network distance becomes too large. However, this problem is not too problematic, as the correct road segments are still included in the resulting path.

A more problematic problem is shown in area B of figure 6.2, where no road segment is selected, while there are candidates that could be in the route. The cause of this issue is if a road network has become too complex (such as the roundabout in area B). Between point 1 and point to in the area, there are five road segments to be traversed. With the map-matching algorithm, only one road segment is added to the path per GPS measurement. While a road segment can be the Hidden Markov State of multiple GPS measurements, this is not the case the other way around. Complex parts of the road network seem to cause severe 'noise' in the segment candidate selection. The map-matching algorithm always tries to create a linked path as much as possible. The linked segment for point 2 in area B is close to and connected to point 3 and its corresponding road segment, instead of being the logical option close to point 2. Newson and Krumm (2009) encountered map-matching problems on complicated parts of their network testing too. They dealt with it by removing what they called the 'junk points' that lead to an incorrect route, but this seems unnecessary for this thesis. While the whole point of map-matching is to not just pick the closest road segment to a GPS measurement, future developments on the method should take the possibility of either selecting the closest segment when the road network becomes too complex, or being able to match multiple road segments to one GPS measurement. While the latter proposition would result in an increased quality of paths of a run, the first would be a more useful addition to the map-matching algorithm, as for this thesis, getting the correct road segment for each GPS measurement is more important than getting a perfect path.

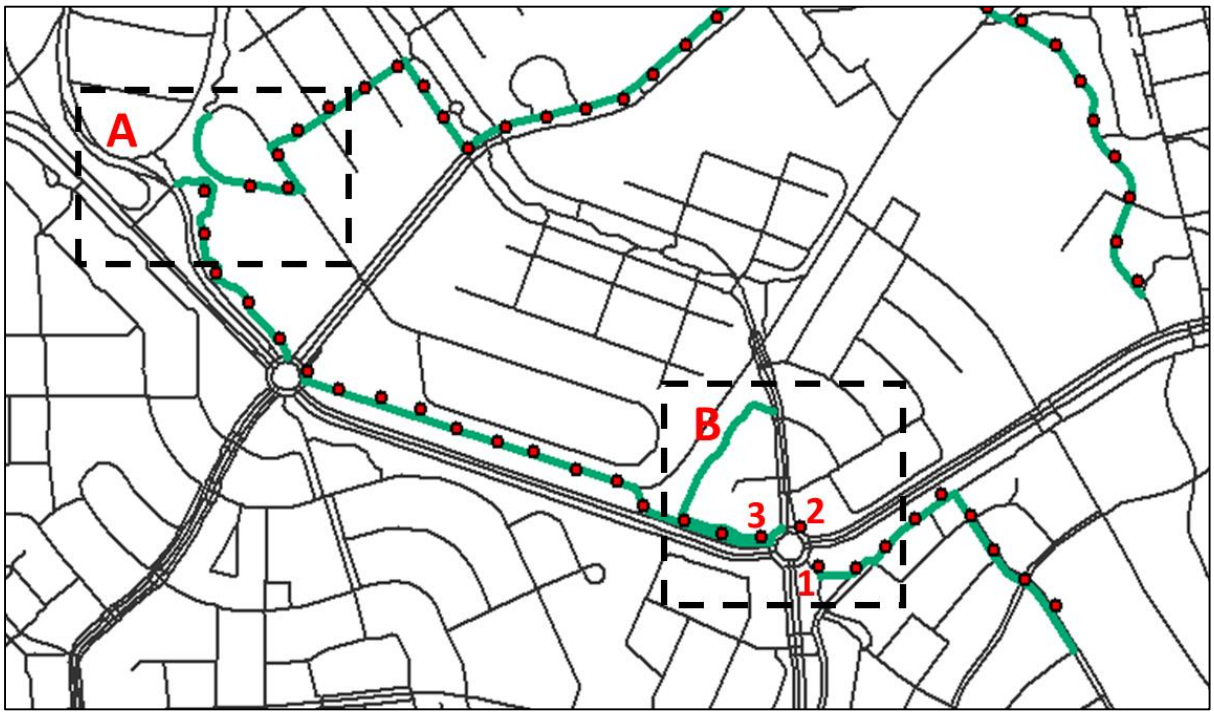


Figure 6.2: Map-matching restrictions for run with ID 'bQKBCvS6GM'

A final problem that can impede with the quality of the map-matching is if the road network does not support the runnable environment at all. A good example of this in the Netherlands, and also in the

research area, is the presence of sandy plains within forests. As can be seen in figure 6.3, a lot of road segments end up at the border of the sandy plain in the Oirschotse Heide, but none actually cross it to connect different ends together. Mitigating the effect of this problem is something that can be a subject of a thesis of its own. Recognizing which areas could be runnable without them being network segments is a complex problem.

Figure 6.3: Inconsistencies of the road network at the Oirschotse Heide

Figure 6.4: Satellite imagery of the Oirschotse Heide (Google, 2019)

6.3 Other Route Enrichment Methods

There are three main methods of assigning the influence values to the runner tracks. Firstly, in some cases it is possible to assign the value based on the matching of information both in the track dataset and the influence sources. Essential is that the influence source features have a unique ID attribute. This attribute is extracted to the GPS measurements in a new attribute. Thereafter it is possible to match these identical number in the two files to extract more information, such as the influence value of the influence source. The code in figure 6.5 shows how this is done for the motorized vehicles influence value.

```
"""
Definition that stores FID ID and distance info to the nearest road segment in the GPS measurements.
"""
def checkMotorizedVehiclesArcPy(network, track, root):
    trackID = os.path.split(track)
    trackID = trackID[1].split('_')
    trackID = trackID[0]
    print trackID
    arcpy.Near_analysis(track, network, 500, "LOCATION")
    #print track
    return track, trackID


"""
Definition that extracts motorized vehicles influence information from road segments to the GPS measurements.
"""
def motorizedVehiclesInfluenceMapping(network, track, root):
    #performance of near analysis and returns the trackfile and the run ID of the track
    f = checkMotorizedVehiclesArcPy(network, track, root)

    #Important to set the encoding, because the 'near' ArcPy method messes it up
    with fiona.open(network, 'r', encoding="utf-8") as segments:
        #print segments.schema

        with fiona.open(f[0], 'r', encoding="utf-8") as points:
            #print points.schema
            schema = points.schema.copy()
            schema['properties']['MV_infl'] = 'float'

            with fiona.open(root+"6TrafficSafety/motorizedVehicles/Output/"+f[1]+"_trackV5.shp", 'w', 'ESRI Shapefile', schema, crs=points.crs) as writer:
                number = 0
                for point in points:
                    for segment in segments:

                        #Match points and segments based on th FID ID given to the segments and points in the ArcPy near functionality
                        if point['properties']['NEAR_FID'] == segment['properties']['FID_ID']:
                            speed = float(segment['properties']['KPH'])
                            pointDist = point['properties']['NEAR_DIST']
                            P = 100.0 - ((100.0 / (speed / 10.0)) * pointDist)
                            if P <= 0.0:
                                P = 0.0
                            influence = P * 1.0
                            #print pointDist, influence
                            point['properties']['MV_infl'] = influence
                            writer.write(point)
                            #print number  kni
                            number += 1

                        else:
                            pass

    #Define the projection to RD_New to be sure
    arcpy.DefineProjection_management(root+"6TrafficSafety/motorizedVehicles/Output/"+f[1]+"_trackV5.shp", arcpy.SpatialReference('RD New'))
```

Figure 6.5: Fetching influence information by matching GPS and source feature IDs

The second method is to just use extract influence information from influence source datasets to the GPS measurements by checking if the geometries overlap one another. To do this, both the GPS measurements and the influence sources are converted into geometries. Thereafter for each point geometry, or GPS measurement, it is checked in which polygon geometry the point is situated, or in the case of working with polyline geometries, it is checked which polylines are within a certain distance of the point. Attribute values from the polyline or polygon geometries and distances between point and influence source can then be extracted to the GPS tracks to be further manipulated into the influence values. Figure 6.6 shows an example of how geometries are matched for the verbal harassment influence factor.

```
1.  def verbalHarassmentInfluenceMapping(root1, root2, VH_influence):
2.      trackList = []
3.      for file in os.listdir(root1):
4.          if file.endswith(".shp"):
5.              #print file
6.              trackList.append(os.path.join(root1, file))
7.      #print trackList
8.      with fiona.open(VH_influence, 'r') as areas:
9.          geometryList = []
10.         for area in areas:
11.             for geometry in area['geometry']['coordinates']:
12.                 #print geometry
13.
14.                 try:
15.                     geom = Polygon(geometry)
16.                 except ValueError:
17.                     for g in geometry:
18.                         geom = Polygon(g)
19.                 info = (area['properties']['VH_infl'], geom, area['properties']['buu
    rtcode'], area['properties']['pointDens'])
20.                 geometryList.append(info)
21.     for file in trackList:
22.         print file
23.         trackID = os.path.split(file)
24.         #trackID = trackID[1]
25.         trackID = trackID[1].split('_')
26.         trackID = trackID[0]
27.         print trackID
28.         #trackID = re.search('Quality/Output/(.+?)_trackV2.shp', file).group(1)
29.         with fiona.open(file, 'r') as track:
30.             schema = track.schema.copy()
31.             schema['properties']['VH_infl'] = 'float'
32.             schema['properties']['buurtCode'] = 'str'
33.             schema['properties']['pointDens'] = 'float'
34.             with fiona.open(root2+"Output/"+trackID+"_trackV3.shp", 'w', 'ESRI Shape
    file', schema, crs=track.crs) as writer:
35.                 for point in track:
36.                     geom = Point(point['geometry']['coordinates'])
37.                     for feat in geometryList:
38.                         if feat[1].contains(geom):
39.                             point['properties']['VH_infl'] = feat[0]
40.                             point['properties']['buurtCode'] = feat[2]
41.                             point['properties']['pointDens'] = feat[3]
42.                             writer.write({'properties':point['properties'],'geometry
    ': mapping(shape(point['geometry']))})
43.                         else:
44.                             pass
45.
46.         arcpy.DefineProjection_management(root2+"/Output/"+trackID+"_trackV3.shp", a
    rcpy.SpatialReference('RD New'))
```

Figure 6.6: Fetching spatial influence information from source features based on overlapping geometries

The third and last method is utilized if the influence sources are stored in a raster file. The Python module of the ArcGIS package has a useful function that can extract the value of a raster cell to a point feature. Figure 6.7 shows the Python code of how this process works. The ArcPy process stores the cell value, which represents the influence per cell, in the runner track in a newly made attribute. To make sure the attribute value is saved correctly, it is also copied to a premade value that gives a more clear indication of the value. Of the three methods, this one is the most standardized. The other two methods are highly dependent on what kind of matching information is available between GPS measurements and influence source features or what data type the influence source features are stored in. Table 6.1 gives an overview of how the GPS measurements are enriched for each influence factor.

```
"""
A definition that finds all shapefiles in a folder and appends them to an empty list
.
"""
def findFiles(root):
    trackList = []
        for file in os.listdir(root+"8SoundPollution/Output/"):
            if file.endswith(".shp"):
                #print file
                trackList.append(os.path.join(root+"8SoundPollution/Output/", file))

        #print trackList
        return trackList


"""
A Definition that extracts the raster cell values of an influence raster to GPS meas
urements in a track.
"""
def getInflValues(root, track, influenceRaster):
    trackID = os.path.split(track)
    trackID = trackID[1].split('_')
    trackID = trackID[0]
    arcpy.ExtractValuesToPoints(track, influenceRaster, root+"9AirPollution/Output/"
+trackID+"_trackV9.shp")


"""
The input variables for the influence mapping.
"""
influenceRaster = root+"9AirPollution/AP_influence.tif"
tracks = findFiles(root)
for track in tracks:
    getInflValues(root, track, influenceRaster)
```

Figure 6.7: Extracting the air pollution raster cell values to the GPS measurements

Enrichment method	Influence factor
Identical information match	Running surface
	Motorized vehicles
Overlapping geometries	Verbal harassment
	Street lighting
	Cyclists
	Variety in surroundings
Raster cell assignment	Natural areas
	Sound pollution
	Air pollution

Table 6.1: Assigning of influence factor to enrichment method

6.4 Influence Aggregation

The method of the influence aggregation has already been discussed in paragraph 3.4. In this paragraph, the consequences of doing it are discussed. A lot of aspects about the spatial influence modeling are not set in stone, due to the exploratory nature of the thesis. An unwanted consequence is that the results themselves might not provide as solid of an understanding of the actual influence on a runner at each moment during the route as might be hoped. The influence aggregation based on the straight line distance between points is a way to give the influence scores on the routes a more realistic representation of reality. While it would have been even more useful to also assess the temporal division between GPS measurements on a route, the lack of such information in the dataset prevents that.

The aggregation process acts as both the standardization of the influences to a scale from -1 to 1 when two-tailed, and for -1 to 0 or 0 to 1 when the influence is one-tailed, as it is to adjust the influences to the route distance they give information on. Due to the unpredictable nature of how the weight for each point will be for all research subjects, the aggregation happens after the standardization and as a result, somewhat abolish the standardization. If a GPS measurement gets a relatively high weight value, as it represents a relatively large distance on the track, and it houses either minimum (-1) or maximum (1) influence values for some of the factors, it is to be expected that those minimum and maximum standardized barriers are crossed when aggregating the influences.

It is argued, however, that this is to be accepted, as it results in more realistic results on the actual influence experience along the route. While it would be possible to not just consider the distance from a GPS measurement to the previous, but to use half of that distance and half of the distance to the next GPS measurement, this would not result in vastly different outcomes, as the same amount of measurements would still need to cover the same amount of distance. The next paragraph explores in greater details the effect of influence aggregation on the descriptive statistics of each influence factor.

6.5 Preliminary Results

When looking at the min and max statistics for the weighted influence values in table 6.2, something that is immediately noticeable is that these values differ significantly from the standardized scale of the influence values. A lot of maximum and minimum values are over five times the standardized limit of -1 and 1, which gives the idea that by weighting the influence values with the distances between points, a lot of outliers have been created, due to significant variations between distances of points. Due to this consideration, it seems useful to not only provide a regression model for the weighted values, but to perform a second analysis on the standardized values, that are always between -1 and 1.

Attribute	Count	Min	Max
Surface quality	17268	-6,37	4,06
Verbal harassment	17268	0,00	2,36
Street lighting	17268	-2,87	3,58
Motorized vehicles	17268	-5,30	0,00
Cyclists	17268	-1,23	0,00
Natural areas	17268	0,00	5,80
Sound pollution	17268	-5,67	3,42
Air pollution	17268	-5,73	0,00
Environmental variety	17268	-5,54	5,64

Table 6.2: Descriptive statistics of weighted influence results on runners

As can be derived from the descriptive statistics of the standardized influence values, as shown in table 6.3, the mean values are more often negative than positive. While a negative minimum is logical for influence factors that scale from -1 to 0, this is also the case for three of the four influence factors that contain both positive and negative values. This suggest an apparent skewed balance in the existence of negative influences over positive influences. This is reflected when looking at the total cumulative influences per GPS measurements. The minimum is lower below 0 influence than the maximum is above 0. The average cumulative influence of a GPS measurement is -1,69.

Attribute	Count	Min	Max	Mean
Surface quality	17268	-0,90	0,80	-0,44
Verbal harassment	17268	-1,00	0,00	-0,23
Street lighting	17268	-1,00	1,00	0,13
Motorized vehicles	17268	-0,99	0,00	-0,15
Cyclists	17268	-0,93	0,00	-0,07
Natural areas	17268	0,00	1,00	0,50
Sound pollution	17268	-1,00	1,00	-0,16
Air pollution	17268	-1,00	-0,54	-0,77
Environmental variety	17268	-1,00	1,00	-0,24
Total influence	17268	-6,60	4,36	-1,69

Table 6.3: Descriptive statistics of standardized influence results on runners

Chapter 7
Model Fitness

7.1 Introduction

In this chapter, the fitness of the methodology is tested with a multiple regression analysis. As mentioned in paragraph 3.5, the regression analysis is used to assess how much of the variance in one dependent variable is explained by multiple independent variables. The 9 influence factors serve as the dependent variables of the analysis. The dependent variable, however, has not been discussed yet. The goal of the analysis is to see if more positive influences lead to more runner activity in an area. The amount of runner activity in an area is therefore taken as the dependent variable.

To calculate the runner activity, the amount of GPS measurements from the 200 runner tracks per square kilometer are used. The areal division for which this is calculated cannot be smaller in scale than the largest geographical context used when modeling the influence factors to avoid problems with the MAUP (see paragraph 2.3). The largest geographical context used is the neighborhood division for the verbal harassment influence factor. This leads to the neighborhood division being used as the areal division for the dependent variable. As only 1% of the total amount of tracks in the original track dataset are used as research subjects, it is useful to assess their coverage among the neighborhoods in the research area. Figure 7.1 shows the neighborhoods in the research area that have at least 1 GPS measurement from one of the 200 used tracks in them. 188 of the 287 neighborhoods in the research area (65,5%) have at least 1 GPS measurement in them. These neighborhoods are therefore included in the regression analysis and its results. Notable is that most neighborhoods on the east and southeast side of the research area are left out of the regression analysis because of this.

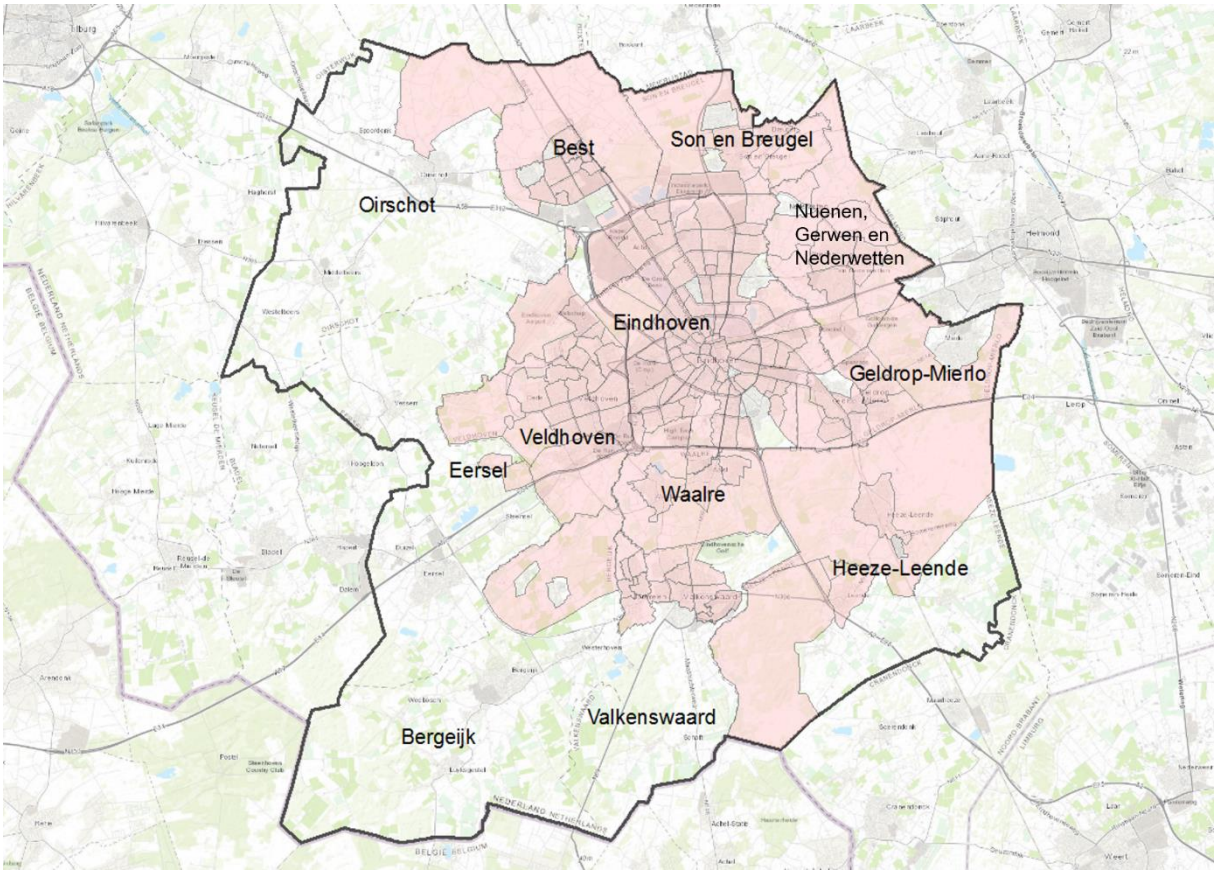


Figure 7.1: Neighborhoods with GPS measurements in them

In the next paragraph, the 6 assumptions of the regression analysis are tested and discussed. Paragraph 7.3 thereafter entails the results of the regression analysis itself.

7.2 Regression Analysis Assumptions

7.2.1 Weighted Influences

1. Scalar variables: All dependent and independent variables are scalar by nature.
2. Influences between dependent and independent variables: As will be discussed in the next chapter, the influences between all the independent variables and the dependent variable are significant.
3. Linearity of the model: The linearity of a regression model when working with multiple regression analyses is difficult to assess. It is, however, possible to deduce this assumption when testing for homoscedasticity. If that assumption is met, the linearity of the model assumption is also met.
4. Homoscedasticity of residues: A scatterplot to test this assumption can be found in appendix 3. The residues do seem to be homoscedastic, but there is a pattern of two groups of standardized residues being higher than the standardized predicted values of the dependent variables, but also noticeably higher than the largest group of standardized residues.

5. Normal distribution of residues: A normality graph and histogram can be found in appendix 3. While the histogram shows a somewhat normal division of residues, there are a lot of outliers, as predicted in paragraph 6.5. This shows in the normality graph, as the residue values do not strictly follow a linear pattern.
6. Autocorrelation between independent variables. In appendix 3, an autocorrelation matrix is included. The important value is the Pearson correlation value, which cannot have a value of below -0,9 or higher than 0,9, as this would suggest two variables explaining the same thing. As can be seen in the correlation matrix, this is not the case and this assumption is met.

7.2.2 Standardized Influences

1. Scalar variables: All dependent and independent variables are scalar by nature.
2. Influences between dependent and independent variables: As will be discussed in the next chapter, the influences between all the independent variables and the dependent variable are significant.
3. Linearity of the model: The linearity of a regression model when working with multiple regression analyses is difficult to assess. It is, however, possible to deduce this assumption when testing for homoscedasticity. If that assumption is met, the linearity of the model assumption is also met.
4. Homoscedasticity of residues: A scatterplot to test this assumption can be found in appendix 4. The residues do seem to be homoscedastic, but there is a pattern of two groups of standardized residues being higher than the standardized predicted values of the dependent variables, but also noticeably higher than the largest group of standardized residues.
5. Normal distribution of residues: A normality graph and histogram can be found in appendix 4. While the histogram shows a somewhat normal division of residues, there are still some outliers to be found in the dataset. As mentioned in the previous assumption there are two groups of outliers that prevent a normal distribution of residues.
6. Autocorrelation between independent variables. In appendix 4, an autocorrelation matrix is included. The important value is the Pearson correlation value, which cannot have a value of below -0,9 or higher than 0,9, as this would suggest two variables explaining the same thing. As can be seen in the correlation matrix, this is not the case and this assumption is met.

7.3 Assessment of Model Fitness

As the assumptions of the regression analysis are tested, the analysis itself can be performed using the SPSS software package. The result of this analysis consists of three output tables, which are included in this paragraph. Just as discussed in paragraph 6.5, both a model with standardized and weighted values, and a model with just the standardized values is discussed.

7.3.1 Weighted Influences

Table 7.1 shows the results of the influence each coefficient, or independent variable, has on the dependent variable and if that influence is significant. The significance is determined through a t-test

for each independent variable, in which it is corrected for the effects of the other independent variables. As can be seen in the table, all influence factors have a significant (lower than 0,05) influence on the dependent variable. This means that the influence of the factors is valid for the entire population. From the B values in table 7.1 the regression equation of the analysis can be deduced. Equation 11 shows the regression equation.

Some initially noticeable results are the direction (B) some of the influences have. Most striking is that a more positive natural influence leads to a decline in the runner activity in a neighborhood. The same can be said for the two other influence factors that show a negative relation with the dependent variable, cyclists and air pollution. In those two cases, however, it can be somewhat explained by respectively the relatively simple modeling method of cyclists and the fact that the air pollution input from Luchtmeetnet (2016) is entirely based on estimations between known sources, measured in select places across the country.

	Unstandardized coefficients		Standardized coefficients	T-test	
Influence Factor	B	Std. Error	Beta	t	Sig.
Constant	70,780	1,923	-	36,798	0,000
Surface Quality	9,615	1,474	0,056	6,522	0,000
Verbal Harassment	36,406	3,197	0,096	11,387	0,000
Street Lighting	20,554	2,086	0,074	9,855	0,000
Motorized Vehicles	16,343	2,608	0,049	6,266	0,000
Cyclists	-48,161	6,734	-0,058	-7,152	0,000
Natural Areas	-18,005	1,901	-0,087	-9,473	0,000
Sound Pollution	28,628	1,293	0,185	22,144	0,000
Air Pollution	-39,600	3,073	-0,134	-12,885	0,000
Environmental Variety	23,951	1,558	0,124	15,372	0,000

Table 7.1: Coefficient table of the weighted influence values

(11) $$Running\ Activity = 70{,}780 + 9{,}615 * Surface\ Quality + 36{,}406 * Verbal\ Harassment + 20{,}554 * Street\ Lighting + 16{,}343 * Motorized\ Vehicles - 48{,}161 * Cyclists - 18{,}005 * Natural\ Areas + 28{,}628 * Sound\ Pollution - 39{,}600 * Air\ Pollution + 23{,}951 * Environmental\ Variety$$

To explain the regression equation a bit more, it has a predicted value of around 71 GPS measurements of runners per square kilometer in a neighborhood if there would be none of the modeled spatial influences effecting the runner. However, this is not a realistic scenario. A runner is prone to experience some of these influences along the run. The aforementioned three influence factors that have negative relation with the dependent variable are the resulting direction of influences that do not match the expected relationship between dependent and independent variable, as it would seem that a worse influence from these factors results in significant higher runner activity in that neighborhood.

The second result table of the regression analysis shows the results of the significance of the entire model, which is shown in table 7.2. This is an ANOVA variance analysis that compares explained and unexplained variances to calculate a model significance. By dividing the sum of squares of the explained and unexplained variances by the degrees of freedom [df], a mean square is calculated. This value entails the mean variance for each degree of freedom. By dividing the explained mean square by the

unexplained mean square an F value can be determined, which in turn leads to the significance of the whole model.

The table immediately gives some worrisome results. The explained Sum of Squares is only a fraction of the total Sum of Squares, indicating a poor model performance. However, the model is still significant. With an F value of 123,423 it can be said with 99% certainty that the independent variables together have a significant influence on how much runner activity there is in a neighborhood. Therefore, the model can be used to assess the exact variation of runner activity that can be explained with the used independent variables.

Variance	Sum of Squares	df	Mean Square	F	Sig.
Explained	12.766.050	9	1.418.450	123,423	0,000
Unexplained	198.338.345	17258	11.493		
Total	211.104.395	17267			

Table 7.2: ANOVA variance analysis of the weighted influence values

The third and final result table of the regression analysis is about the fitness of the used model and in this case, methodology. This table is shown in table 7.3. The R value in the table signifies the correlation between the measures dependent values (amount of GPS measurements per square kilometer in a neighborhood) and the predicted dependent variables, which entail the runner activity in a neighborhood when looking at the results of the influence factors.

The R Square value is the square of the R value and comprehends the percentage of explained variance in the dependent variable by the independent variables. The adjusted R Square value signifies the fitness of the model when using a multiple regression analysis, as it adjusts the R square value based on the amount of independent variables in the regression model.

As expected from the ANOVA table, the performance of the model is quite poor. With an adjusted R of only 0,060, the weighted model only seems to explain 6,0% of the variance in the runner activity. As discussed in paragraph 6.5, the weighting of the influence values might have had a negative impact on the model outcomes, as it seems to severely create outliers.

R	R Square	Adjusted R Square	Std. Error of the Estimate
0,246	0,060	0,060	107,203

Table 7.3: Model performance of the weighted influence values

7.3.2 Standardized Influences

When only standardizing the spatial influences on a scale of -1 to 1, the results seem to be completely different. When looking at the coefficient table (7.4), the effect each influence factor has on the activity of runners in a neighborhood is much bigger than was the case with the weighted values. While it might seem strange that each influence factor has such a large impact on the runner activity in a neighborhood, it is actually due to the standardized nature of the influence values. The B value from the

table indicates a change in activity with a step of 1,0 of the influence value. As the scale of a factor is at most 2,0 wide, getting such a big change in an influence factor in a neighborhood seems quite consequential.

	Unstandardized coefficients		Standardized coefficients	T-test	
Influence Factor	B	Std. Error	Beta	t	Sig.
Constant	-125,445	11,561	-	-10,851	0,000
Surface Quality	12,031	1,660	0,058	7,246	0,000
Verbal Harassment	-21,396	3,504	-0,054	-6,106	0,000
Street Lighting	20,844	2,131	0,073	9,783	0,000
Motorized Vehicles	12,532	2,935	0,032	4,269	0,000
Cyclists	-51,998	6,952	-0,060	-7,480	0,000
Natural Areas	-15,188	2,138	-0,056	-7,103	0,000
Sound Pollution	49,038	1,607	0,275	30,506	0,000
Air Pollution	-304,457	15,385	-0,187	-19,789	0,000
Environmental Variety	27,231	1,720	0,123	15,831	0,000

Table 7.4: Coefficient table of the standardized influence values

Differences in model performance are also very large when looking at the ANOVA table and the model fit table (table 7.5 and table 7.6). Without weighting the influence values, the performance of the model is significantly better. Both the independent variables and the model itself are still significant and the ratio of explained and unexplained variance is better. Model performance itself shows an R value of 0,298, which indicates an average relationship between the dependent variable and the independent variables. The corresponding explained variance (adjusted R square) is 8,8%. This means that 8,8% of variation in runner activity in a neighborhood is explained with the in this thesis used standardized spatial influence factors and spatial influence modeling methods.

Variance	Sum of Squares	df	Mean Square	F	Sig.
Explained	18.766.471	9	2.085.163	187,096	0,000
Unexplained	192.337.924	17258	11.144		
Total	211.104.295	17267			

Table 7.5: ANOVA variance analysis of the standardized influence values

R	R Square	Adjusted R Square	Std. Error of the Estimate
0,298	0,089	0,088	105,57

Table 7.6: Model performance of the standardized influence values

The multiple regression models have both performed poorly. Paragraph 8.2 will go into more detail about both the weaknesses of the research methodology and the regression model. Before that, however, a quick test is done to check the influence of adding an additional independent variable to the regression analysis. A byproduct of modeling the cyclists influence factor is the population density of

the neighborhoods. As this is a scalar variable that is measured at the exact spatial scale of the dependent variable, it can be used to provide some additional insights in the workings of the model. When adding the population density to the regression analysis the explained variation in runner activity per neighborhood increases from 8.8% to 11,2%. This is not meant a creation of new results, but just to show how such a factor that did not derive from the scientific literature study has a relative big effect on the regression analysis. Utilizing population density, however could be related willingness of runners to travel somewhere to start a run instead of doing it from their place of living (Šilerytė, 2015).

Chapter 8
Conclusion

8.1 Answering the Research Questions

In paragraph 1.4 of the research introduction, three main questions were formulated in correspondence to four research objectives. The entirety of this master thesis since then serves to answer these research questions. While for a complete understanding of the answers to these questions it is better to explore the previous chapters, this paragraph serves as a short answer to the research questions.

RQ1: Which spatial factors in the running environment have a positive influence on the activeness of runners and which factors have a negative influence on it?

The first research question focused on the scientific literature related to the subject in the scientific field of PABE. With the help of the paper by Ettema (2015) and three requirements stated by Allen-Collinson (2008), which are maintaining momentum, enhancing performance, and avoiding injury, sixteen different spatial influence factors for runners were determined. These factors and the kind of influence the scientific literature states them to have can be found in tables 2.1. For a clear overview on them, the factors were divided into five categories:

1. Surface Quality: The fit of the running surface for runners. What kind of surfaces benefit the activeness of runners and helps them avoid injury and maintain momentum and what kind of surfaces lack the positive influence.
2. Social Safety: Runners are subjected to social interactions during their run. Depending on how these envelop, the runner can be motivated or demotivated by them.
3. Traffic Safety: When running in a dense country such as the Netherlands, but applicable to all countries, is the amount of interactions a runner has with other traffic. For the Netherlands, this category is divided into motorized vehicles, cyclists and pedestrians. The latter, however, is not taken into account in the analysis as it is too complicated to accurately measure and analyze.
4. Surrounding Environment: When modeling the spatial influence on runners, it is logical to include the physical surroundings, which this category entails.
5. Route Information: The fifth category is a minor one compared to the others, as it only encompasses the amount of information presented to the runner during or in advance of their

run. As this is not too widespread in the Netherlands, it is chosen to not include this category in the analyses.

The spatial influence factors in the five categories provide the input information for the analyses done in this thesis and thus also for the second and third research question.

RQ2: How can the spatial influence factors be operationalized with the available geographical data?

To operationalize the influence factors, three different approaches to model spatial influences have been discussed and used; the surface based, distance based and areal influence modeling methods. While it was initially assumed that both the influence factor itself as the data used to model it had a significant influence on which method was used. As became clear when modeling the influences, predominantly the utilized data determined the modeling method. The most important cause is that most influence factors do not give an immediate clarity in how to best approach their modeling or when they do, suitable data cannot always be found. While surface quality suggests aspects of the ground the runner is running on and therefore the surface based approach, most other factors were less clear. A good example of this is the street lighting influence factor. The influence factor would suggest either a distance based or areal modeling method, as the influence of street lighting on a runner seems dependent on the distance a runner is away from a light source. As only a part of the research area had data available on the exact location of light sources, other data was utilized. A network dataset with information on how well lid the road segments were was used, and therefore the data determined the modeling method as surface based. In the end, nine influence factors are modeled in this thesis with one of the three modeling methods (see chapter 5).

Besides the spatial influence modeling methods, different ways to enrich the tracks were also used. These methods were completely dependent on the data for each influence factor, and did not turn out to be linked to either of the three spatial modeling methods. To fetch the relevant influence information from influence source features to the GPS measurements, three methods were utilized:

1. In some cases, the influence source features contained unique identifiers that were easily extracted to the GPS measurements. This resulted in the matching of the right feature to the right measurement and the exchange of information between them.
2. The second method is to use the geometries of the GPS measurements and the influence source features as a way of determining the distances between them and calculating the influence with that distance, or to check if the geometries were overlapping and when so extracting the necessary information from them.
3. The last enrichment method concerned the influence factors that utilized an influence raster, instead of vector features. In these cases the influence raster cell values can be extracted to the GPS measurements that are positioned on top of them.

RQ3: To what extent can the operationalized spatial influences and research methodology be validated?

Validation of the research results has been done by means of statistical testing. As analysis method, a multiple regression was chosen, as this method would result in an estimated fitness of the research methodology. When aggregating the influence values for the factors, it became apparent that the

method of aggregation, through distances between points, led to a large number of outliers. To assess the effect did would have on the regression analysis, a second model was executed wherein the influence values would be just standardized, and not weighted. While this second model did perform better than the first, with de independent variables explaining 8,8% of the variation in runner activity over 6,0% of the weighted model, the performance leaves a lot to be desired and to be improved

8.2 Discussion & Recommendation for Further Research

The results gathered in this master thesis are not perfect or to not be disputed. As there is not a lot of precedence in the modeling of spatial influences on runners, it remains to be seen if the used research methodology can be used for other research, bot scientific and non-scientific. In this paragraph, I will discuss four aspects of the research methodology and results that can be improved on that can lead to better model performances when using the multiple regression analysis for validation.

8.2.1 Influence Functions

The first to be improved aspect of the thesis is the used influence functions when modeling the influence factors. These functions are mentioned in chapter 5 and can be found in the Python scripts in the accompanying data files. All of these functions have been estimated by the researcher, without a lot of theoretical background. This was necessary due to the lack of prior research into the spatial influences on runners combined with the usage of GIS.

It would, however, have been advisable to conduct a more logical reasoning when determining the influence functions for the factors. I too would have helped to play around some more with different influence functions to assess the affect it would have on the overall results. Overall time and hardware limitations did not invite this experimenting, as changing a lot could result in many hours of processing time. However, this is no excuse and the results would have a more grounded validity when the experimenting would have happened.

Key in achieving this, besides experimenting with different influence functions, is to gather quantitative data to support the methods. Interviews with either experts on physical activity in the build environment or frequent runners could improve the insight into how the influence factors need to be modeled greatly.

8.2.2 Input Tracks

A second aspect of the research methodology that could be improved is how many tracks are used to as input for the methods. For this thesis only 200 of the available 21405 tracks were utilized. Using such a small sample resulted in not all parts in the research are being included. In addition, it could have resulted in more randomized results, as the randomly chosen runs could have been unrealistically clustered in some areas, while being strangely absent in others. To decide if the used methods and influence factors have been the correct combinations, a larger pool of research tracks could help in deciding this.

While a total of 17.268 GPS measurements in 200 tracks were used in this analysis, this does not necessarily lead to a lot of data per neighborhood. As discussed in paragraph 7.1 one third of the neighborhoods in the research area are not included in the regression analysis, because of having no GPS measurements in them. For most other neighborhoods, the question arises if 'only' having 17.268 GPS measurements spread out over them is enough to validate the methodology. If possible, increasing the pool of random selected tracks from the complete GeoJSON file until the coverage of each neighborhood is significantly increased might be a good idea.

As an illustration, figure 8.1 shows two areas in the city of Eindhoven with the GPS tracks of the complete dataset in blue and the randomly chosen tracks in red. The left area is in the neighborhood called 'Eliasterrein-Vonderkwartier', next to the inner city of Eindhoven and close to its main railway station. The coverage of the random selected tracks of the total amount of tracks in that area is minimal. On the right, part of the 'Genneper Parken', a city park, on the south side of the city of Eindhoven is shown. The coverage in this area seems to be significantly better.

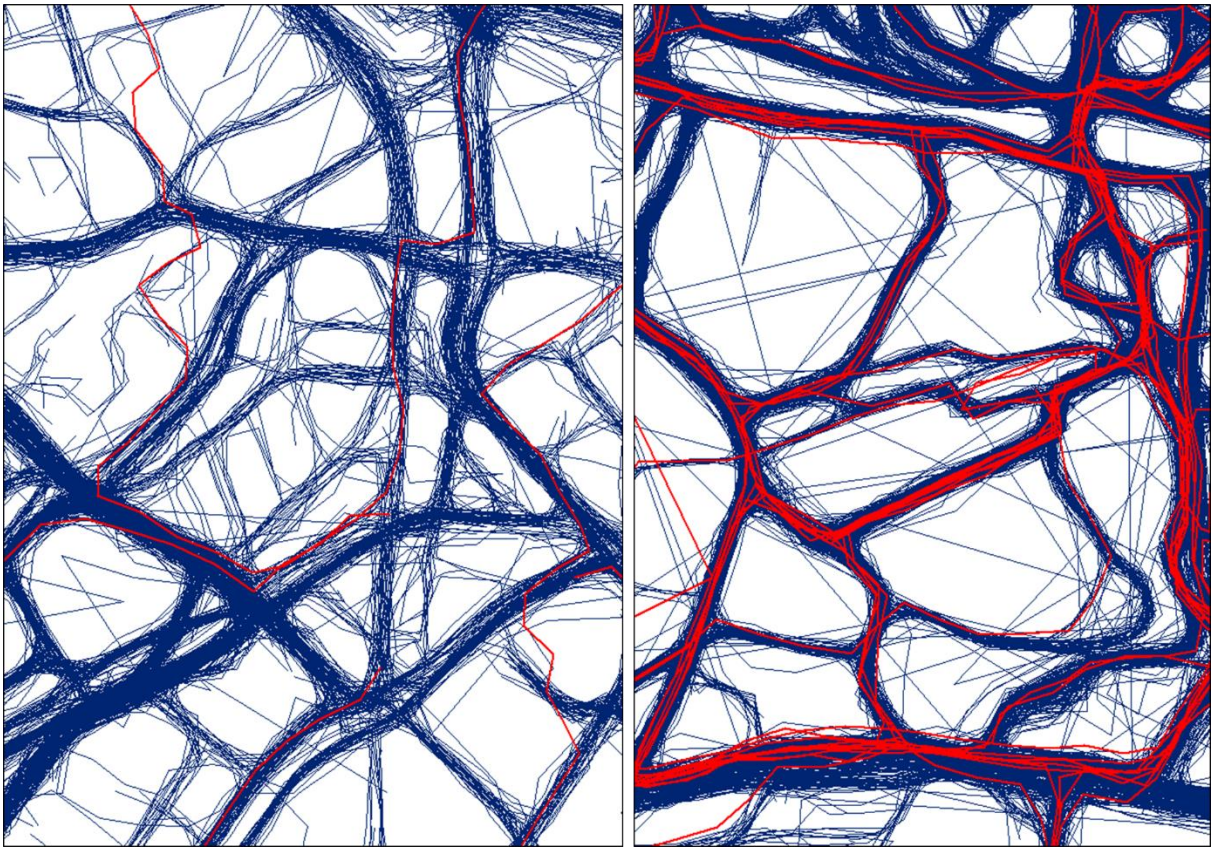


Figure 8.1: All the runner tracks (blue) and the used tracks (red) in the Elias-Vonderkwartier (left) and Genneper Parken (right)

The above example shows that just using the 200 randomly selected track for analysis, does not provide a clear picture of where people run in and around Eindhoven. A larger pool of selected runner tracks is necessary for further research.

8.2.3 Chosen Dependent Variable

Another reason that can cause the poor regression model performance is the chosen dependent variable. In this thesis the amount of GPS points per square kilometer per neighborhood was chosen as the dependent variable. It could be that another variable needs to be chosen to use as the dependent one. An example could be to modify the current one by also accounting for the amount of runnable network there is in each neighborhood. As mentioned before, no complete runnable network could be generated for this research, as there are both runnable areas that are not represented as network segments and an inherent behavior of runners to run off-road and take shortcuts. A good way that might be mitigated is to use runners flows to determine road segments where they are lacking in the data (see figure 8.1). This would be prone to unreliability, however. Besides this example, other indications of runner activity in certain spatial units can be found.

8.2.3 Chosen Factors

Finally, the biggest improvement to be made, is the plethora of different influence factors used in this thesis. As evident from chapters 2 and 3, more influence factors were already gathered from the scientific literature than have been utilized in the analyses. Especially the influence factor 'closeness to stimulating environments' could have been of great value to the methodology. It was discarded, due to the lack of information on the research route subjects and where they lived or how far they travelled to start their route. As running is a recreational physical activity, it can be interesting to study how far runners are willing to travel from their home or work to the starting point of the run. As Šilerytė (2015, p. 96) in a similar master thesis pointed out:

"…the Runability Index has an extra challenge of explaining 'whether or not' people are likely to use the certain space for recreation."

This sheds some light on the fact that while the psychological aspect of runners' decisions to run routes have been ignored in this thesis, due to the lack of data, it would be wise to include it when researching the topic in the future.

In addition, the directions of the influence values discussed in paragraph 7.3 also seems to tell that a factor that is not included in the analysis that is related to the psychological aspect also played a big role in the regression analysis. This is the population density of an area. Both verbal harassment and nature as an influence factor showed a negative relation with the amount of runner activity in an area. As discussed earlier in this thesis, verbal harassment is more likely to occur in densely populated areas and natural aspects of the surroundings decline when the 'build' part of the surroundings, and corresponding the population density, increases. This could be an indication that population density is an influence factor that needs to be discussed in further research.

8.2.5 Practical Limitations

For the last aspect of the thesis that is discussed, a recommendation is to tackle this research methodology without the monetary constraints present in this master thesis. This could mean that all determined spatial influence factors can be modeled with the most suitable data for it. In this thesis, for

example, part of the noise pollution factor and the air quality factor were difficult to model, as the most suitable data for it was owned by a commercial company and not available as open data. When these datasets could be acquired and utilized, however, it will benefit the completeness and precision of the research methodology outcomes greatly.

A third and final recommendation for further research is to rewrite the scripts used for the research methodology to use exclusively open source software. As evident from this master thesis, the ArcGIS coupled python package ArcPy was used extensively throughout the research. However, it is possible to work around using ArcPy with the help of open source python libraries such as OGR, GDAL, Shapely, PyShp and Fiona. While already utilized extensively, their contribution to the research methodology can be expanded. It remains to be seen, however, if this has any additional benefits besides using exclusively open source software.

This thesis has dealt with 2D data only, but it is feasible to include datasets with a height component in the future. If these datasets are included, it is entirely reasonable that other methods of spatial influence modeling are more effective than the ones used in this research, as they are constructed around 2D datasets. A method this could be done with is to use a visibility component. For example, if an influence source is close but the view on it is obstructed by something else, it is logical its influence is different than if the view on it was not obstructed or that it is even non-existent. Influence probability therefore does not only degrade over distance, but also based on what is in between runner and influence source. Therefore, integrating a 3D component could lead to using different influence factors or adding them to the current selection, as the openness of a space could be integrated, as runners might prefer of avoid runs wherein they can get a view.

8.3 Reflection

This master thesis stems from a combination of an increased interest in studying physical activity in the build environment and the increased integration of geographical information in these studies. While this trend is still on the rise, by now a lot has been researched and discussed regarding this topic. A striking hole in this discussion, however the subject of a runner. Scientific research still has its focus on either cyclists or pedestrians, while the 'middle ground' runner is lacking in studies. This leads in turn to both a problem and an interest. The problem is the apparent lack of preceding studies into the behavior of runners in the build environment in combination with geographical information. The interest is that there is always a curiosity for that which is still quite unknown.

In this thesis, runner training route information, gathered by the Technical University of Eindhoven and Fontys Sporthogeschool Eindhoven, is used in combination with through scientific literature determined influence factors to assess the spatial influences a runner experiences during the run and to what extent. The master thesis is therefore a combination of research into physical activity and the inclusion of geographical information. By using multiple different spatial influence methods, GIS software and statistical analysis, the use of the gathered influence factors and chosen methods has been tested and discussed.

Because of the explorative nature of this research into using new methods for modeling spatial influences on runners, this research underwent its fair share of holdups and delays. Geographical data

can sometimes be incredibly slow to work with, due to both its data size, inefficiencies in accessing it and limitations of the software and hardware, which made testing new methods on a suitable size of runner tracks problematic. In addition, the available geographical data is not always suitable to perform all the goals and tasks that were stated at the start of the research process. Finishing this master thesis, however, did prove a great learning tool.

Bibliography

Addy, C.L., Wilson, D.K., Kirtland, K.A., Ainsworth, B.E., Sharpe, P. & Kimsey, D. (2004). Associations of Percieved Social and Physical Environmental Supports With Physical Activity and Walking Behavior. *American Journal of Public Health* 94 (3), pp. 440-443.

Allen-Collinson, J. & Hockey, J. (2007). 'Working Out' Identity: Distance Runners and the Management of Disrupted Identity. *Leisure Studies* 26 (4), pp. 381-398. DOI: 10.1080/02614360601053384

Allen-Collinson, J. & Hockey, J. (2013). From a Certain Point of View: Sensory Phenomenological Envisionings of Running Space and Place. *Journal of Contemporary Ethnography* 10 (5), pp. 1-21. DOI: 10.1177/0891241613505866

ArcGIS (n.d.). What is ArcPy? Retrieved 16 November 2016 from: http://pro.arcgis.com/en/pro-app/arcpy/get-started/what-is-arcpy-.htm

Atlas Leefomgeving (2018a), Kaarten. Retrieved on 17 December 2018 from: https://www.atlasleefomgeving.nl/kaarten

Atlas Leefomgeving (2018b), Geluid. Retrieved on 15 December 2018 from: https://www.atlasleefomgeving.nl/meer-weten/geluid

Austin, P.C., D. van Klaveren, Y. Vergouwe, D. Nieboer, D.S. Lee & E.W. Steyerberg (2016). Geographic and temporal validity of prediction models: Different approaches were useful to examine model performance. Journal of Clinical Epidemiology 79, pp. 76-85. DOI: 10.1016/j.jclinepi.2016.05.007

Basta, L.A., Richmond, T.S. & Wiebe, D.J. (2010). Neighborhoods, daily activities, and measuring health risks experienced in urban environments. *Social Science & Medicine* 71 (11), pp. 1943-1950. DOI: 10.1016/j.socscimed.2010.09.008

Bernaards, C.M., Hildebrandt, V.H. & Stubbe, J.H. (2011). Trendrapport Bewegen en Gezondheid 2010/2011. Leiden: TNO.

Bernaards, C.M., Hildebrandt, V.H. & Hofstetter, H. (2015). Trendrapport Bewegen en Gezondheid 2000/2014. Leiden: TNO.

Bodin, M. & Hartig, T. (2003). Does the outdoor environment matter for psychological restoration gained through running? *Psychology of Sport and Exercise* 4, pp. 141-153. DOI: 10.1016/S1469-0292(01)00038-3

Bona, P. (2000). Accuracy of GPS Phase and Code Observations in Practice. *Acta Geodaetica et Geophysica Hungarica* 35 (4), pp. 433-451.

Breuer, C., Hallmann, K. & Wicker, P. (2011). Determinants of sport participation in different sports. *Managing Leisure* 16 (4), pp. 269-286. DOI: 10.1080/13606719.2011.613625

Brunyé, T.T., Gagnon, S.A., Gardony, A.L., Gopal, N., Holmes, A., Taylor, H.A. & Tenbrink, T. (2015). Where did it come from, where do you go? Direction sources influence navigation decisions

during spatial uncertainty. The Quarterly Journal of Experimental Psychology 68 (3), pp. 585-607. DOI: 10.1080/17470218.2014.963131

Bureau-Maris (n.d.). N305 Waterlandseweg Almere. Retrieved 1 December 2016 from: http://www.bureau-maris.nl/projecten/n305-waterlandseweg-almere/

Butler, H., Daly, M., Doyle, A., Gillies, S., Schaub, T. & Schmidt, C. (2008). GeoJSON. Retrieved 4 May 2017 from: http://geojson.org/geojson-spec.html

Cebrecos, A., Díez, J., Gullón, P., Bilal, U., Franco, M. & Escobar, F. (2016). Characterizing physical activity and food urban environments: a GIS-based multicomponent proposal. *International Journal of Health Geographics* 15 (1), pp. 1-13. DOI: 10.1186/s12942-016-0065-5

Chaix, B. (2009). Geographic Life Environments and Coronary Heart Disease: A Literature Review, Theoretical Contributions, Methodological Updates and a Research Agenda. *Annual Review of Public Health* 30, pp. 81-105. DOI: 10.1146/annurev.publhealth.031308.100158

Chimenti, L., Morici, G., Paterno, A., Bonanno, A., Vultaggio, M., Belliam V. & Bonsignore, M.R. (2009). Environmental conditions, air pollutants, and airway cells in runners: a longitudinal field study. Journal of Sports Science 27 (9), pp. 925-935. DOI: 10.1080/02640410902946493

Clark, S. (2015). Running into trouble: constructions of danger and risk in girls' access to outdoor space and physical activity. *Sport, Education and Society* 20 (8), pp. 1012-1028. DOI: 10.1080/13573322.2013.866548

Comber, A., C. Brunsdon & E. Green (2008), Using a GIS-based network analysis to determine urban greenspace accessibility for different ethnic and religious groups. *Landscape and Urban Planning* 86, pp. 103-114. DOI: 10.1016/j.landurbplan.2008.01.002

Cook, R.D. & S. Weisberg (1982). Criticism and Influence Analysis in Regression. Sociological Methodology 13, pp. 313-361. DOI: 10.2307/270724

Council for Regulatory Environmental Modeling (2009). Guidance on the Development, Evaluation, and Application of Environmental Models. DC: U.S. Environmental Protection Agency.

Cova, T.J., Miller, H.J., Beard, K., Frank, A.U. & Goodchild, M.F. (2008). Geographic Information Science. Berlin: Springer-Verlag.

Dark, S.J. & Bram, D. (2007). The modifiable areal unit problem (MAUP) in physical geography. *Progress in Physical Geography* 31 (5), pp. 471-479. DOI: 10.1177/0309133307083294

Denuit, M., M. Mesfioui & J. Trufin (2019). Bounds on Concordance-Based Validation Statistics in Regression Models for Binary Responses. Methodology and Computing in Applied Probability 21, pp. 491-509. DOI: 10.1007/s11009-017-9613-0

Diez-Roux, A.V. & Mair, C. (2010). Neighborhoods and health. *Annals of the New York Academy of Sciences* 1186, pp. 125-145. DOI: 10.1111/j.1749-6632.2009.05333.x

El Helou, N., Tafflet, M., Berthelot, G., Tolaini, J., Marc, A., Guillaume, M., Hausswirth, C. & Toussaint, J.F. (2012). Impact of Environmental Parameters on Marathon Running Performance. PLoS ONE 7 (5), pp. 1-9. DOI: 10.1371/journal.pone.0037407

Emam, K.E., A. Brown & P. AbdelMalik (2009). Evaluating Predictors of Geographic Area Population Size Cut-offs to Manage Re-identification Risk. Journal of the American Medical Informatics Association 16 (2), pp. 256-266. DOI: 10.1197/jamia.M2902

Esri (n.d.). ArcGIS Overview. Retrieved 16 November 2016 from: http://www.esri.com/software/arcgis

Ettema, D. (2015). Runnable Cities: How Does the Running Environment Influence Perceived Attractiveness, Restorativeness and Running Frequency? *Environment and Behavior*, pp. 1-21. DOI: 10.1177/0013916515596364

Ettema, D. & Smajic, I. (2015). Walking, Places and Wellbeing. *The Geographical Journal* 181 (2), pp. 102-109. DOI: 10.1111/geoj.12065

Falt, E. (2006). Sport and the Environment. *Environmental Health Perspectives* 114 (5), pp. 268-269.

Fatfouta, R., Schulreich, S., Meshi, D. & Heekeren H. (2015). So Close to a Deal: Spatial-Distance Cues Influence Economic Decision-Making in a Social Context. PLoS ONE, pp. 1-9. DOI: 10.1371/journal.pone.0135968

Forney Jr., G.D. (2005). The Viterbi Algorithm: A Personal History. Retrieved 19 November 2016 from: https://arxiv.org/abs/cs/0504020v2

Forsyth, A. (2000). Analyzing Public Space at a Metropolitan Scale: Notes on the Potential for Using GIS. *Urban Geography* 21 (2), pp. 121-147. DOI: 10.2747/0272-3638.21.2.121

Geisberger, R., Sanders, P., Schultes, D. & Delling, D. (2008). Contraction hierarchies: faster and simpler hierarchical routing in road networks. Berlin: Springer-Verlag.

Gladwell, V.F., Brown, D.K., Wood, C., Sandercock, G.R. & Barton, J.L. (2013). The great outdoors: how a green exercise environment can benefit all. *Extreme Physiology & Medicine* 2 (3), pp. 1-7. DOI: 10.1186/2046-7648-2-3

Gonzalez, H., Han, J., Li, X., Myslinskam, M. & Sondag, J.P. (2007). Adaptive faster path computation on a road network: a traffic mining approach. In: *VLDB 2007*. Vienna: VLDB Endowment.

González-Bailón, S. (2006). Big data and the fabric of human geography. Dialogues in Human Geography 3 (3), pp. 292-296. DOI: 10.1177/2F2043820613515379

Goodchild, M.F. (2010). Twenty years of progress: GIScience in 2010. Journal of Spatial Information Science 1, pp. 3-20. DOI: 10.5311/JOSIS.2010.1.2

Groenink, J.W. (2013). Hardlopen in de openbare ruimte. Utrecht, The Netherlands: Utrecht University.

Harris, J.K., Lecy, J., Hipp, J.A., Brownson, R.C. & Parra, D.C. (2013). Mapping the development of research on physical activity and the built environment. *Preventive Medicine* 57, pp. 533-540. DOI: http://dx.doi.org/10.1016/j.ypmed.2013.07.005

Heywood. I., Cornelius, S. & Carver, S. (2011). An introduction to geographical information systems. Harlow/England: Pearson Education Limited. Fourth Edition.

Hockey, J. & Allen-Collinson, J. (2006). Seeing the way: visual sociology and the distance runner's perspective. *Visual Studies* 21 (1), pp. 70-81. DOI: 10.1080/14725860600613253

Hover, P. (2013). Derde loopgolf dient zich aan. Retrieved 11 October 2016 from: http://www.sportnext.nl/berichten/derde_loopgolf_dient_zich_aan

Johnston, C.A.M., Taunton, J.E., Lloyd-Smith, D.R. & McKenzie, D.C. (2003). Preventing running injuries: Practical approach for family doctors. *Canadian Family Physician* 49 (9), pp. 1101-1109.

Jongegeel-Grimen, B., Busschers, W., Droomers, M., van Oers, H.A.M., Stronks, K. & Kunst, A.E. (2013). Change in Neighorhood Traffic Safety: Does It Matter in Terms of Physical Activity? *PloS ONE* 8 (5), pp. 1-12. DOI: 10.1371/journal.pone.0062525

Kluitenberg, B., Middelkoop, M. van, Diercks, R.L., Hartgens, F., Verhagen, E., Smits, D.W., Buist, I. & Worp, H. van der (2013). The Nlstart2run study: health effects of a running promotion program in novice runners, design of a prospective cohort study. *BMC Public Health* 13, pp. 1-7. DOI: 10.1186/1471-2458-13-685

Kwan, M.P. (2012). The Uncertain Geographic Context Problem. *Annals of the Association of American Geographers* 102 (5), pp. 958-968. DOI: 10.1080/00045608.2012.687349

Lee, C. & Moudon, A.V. (2008). Neighbourhood design and physical activity. *Building Research & Information* 36 (5), pp. 395-411. , DOI: 10.1080/0961321080204554

Loureiro, A. & Veloso, S. (2014). Outdoor Exercise, Well-Being and Connectedness to Nature. *Psico* 45 (3), pp. 299-304.

Luchtmeetnet (2016). Luchtkwaliteitsindex Nederland. Retrieved 2 December 2016 from: https://www.luchtmeetnet.nl/kaart/noord-brabant

Luchtmeetnet (2019), Uitleg. Retrieved on 11 February 2019 from: https://www.luchtmeetnet.nl/uitleg

Mardikyan, S. & O.S. Darcan (2006). A Software Tool for Regression Analysis and its Assumptions. *Information Technology Journal* 5 (5), pp. 884-891. DOI: 10.3923/itj.2006.884.891

McCloy, K. (2006), Remote sensing, GIS and modelling. Boca Ranton, USA: Taylor & Francis Group. Second Edition.

McKenzie, D.C. & Boulet, L.P. (2008). Asthma, outdoor air quality and the Olympic Games. *Canadian Medical Association Journal* 179 (6), pp. 543-548. DOI: 10.1503/cmaj.080982

Menon, C. (2012). The bright side of MAUP: Defining new measures of industrial agglomeration. *Papers in Regional Science* 91 (1), pp. 3-28. DOI: 10.1111/j.1435-5957.2011.00350.x

Municipality of Eindhoven (2015). Bijlage concept-sportvisie. Retrieved 17 October 2016 from: http://eindhoven.notudoc.nl/cgi-bin/showdoc.cgi/action=view/id=1411189/type=pdf/Bijlage_concept-sportvisie.pdf

Newson, P. & Krumm, J. (2009). Hidden Markov Map Matching Through Noise and Sparseness. Retrieved 14 October 2016 from: http://research.microsoft.com/en-us/um/people/jckrumm/Publications%202009/map%20matching%20ACM%20GIS%20camera%20ready.pdf

Openshaw, S. & Taylor, P.J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In Wrigley, N. (Ed.), *Statistical applications in spatial sciences* (pp. 127-144). London: Pion.

Ostermann, F.O. & Timpf, S. (2009). Use and appropriation of space in urban public parks. *Geographica Helvetica* 64 (1), pp. 30-36. DOI: 10.5194/gh-64-30-2009

Piwowar, J. (2012). GI systems and science: network analysis. Regina/Canada: University of Regina, Department of Geography. Retrieved 4 November 2016 from: http://uregina.ca/piwowarj/geog303/12%20-%20Network%20Analysis.pdf

Van Poortvliet, W. (n.d.). De Sloterplas. Retrieved 20 November 2016 from: https://hogguerstraat887.wordpress.com/de-omgeving/recreatie-en-sport/de-sloterplas/

Python (n.d.). About Python. Retrieved 16 November 2016 from: https://www.python.org/about/

Reiling, M. & T. Dolders (2015). *Running Amsterdam: Designing a runner friendly city* (Master's Thesis, Wageningen University and Research Centre, Wageningen, The Netherlands). Retrieved from https://library.wur.nl/WebQuery/theses

Rodríguez-Puente, R. & Lazo-Cortés, M.S. (2013). Algorithm for shortest path search in Geographic Information Systems by using reduced graphs. *Springer Plus* 2, p. 1-13. DOI: 10.1186/2193-1801-2-291

Saltelli, A. & Annoni, P. (2010). How to avoid a perfunctory sensitivity analysis. *Environmental Modelling & Software* 25 (12), pp. 1508-1517. DOI: 10.1016/j.envsoft.2010.04.012

Scheerder, J. & Breedveld, K. (2015). Running Across Europe, The Rise and Size of One of the Largest Sport Markets. Basingstoke: Palgrave Macmillan.

Shephard, R.J. (1984). Athletic performance and urban air pollution. Canada Medical Association Journal 131 (2), pp. 105-109.

Shipway, R. & Holloway, I. (2010). Running free: Embracing a healthy lifestyle through distance running. *Perspectives in Public Health* 130 (6), pp. 270-276. DOI: 10.1177/1757913910379191

Šilerytė, R. (2015). *Analysis of Urban Space Networks for Recreational Purposes based on Mobile Sports Tracking Application Data* (Master's Thesis, Technical University of Delft, Delft, The Netherlands). Retrieved from https://repository.tudelft.nl/islandora/object/uuid%3Ac04b441c-7e66-4a72-ba6c-423beac4e17f

Spotzi (2016). Geluidshinder Nederland. Retrieved 2 December 2016 from: http://spotzi.com/nl/kaarten/milieu-en-omgeving/geluid-en-luchtkwaliteit/geluidshinder-nederland/

Statistics Netherlands (2011). Actieve en passieve sportparticipatie; personen van 6 jaar en ouder. Retrieved 11 October 2016 from: http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=80909NED&D1=11&D2=a&D3=0&D4=a&VW=T

Tahoe Transportation (2015). TRPA Governing Board Unanimously Approved the TTD's SR89/Fanny Bridge Community Revitalization Project. Retrieved 1 December 2016 from: http://tahoetransportation.org/blog/page/2/

Veiligheid.nl (2014). Meer sportblessures door beginnende hardlopers. Retrieved 11 October 2016 from: http://voorkomblessures.veiligheid.nl/nieuws/meer-sportblessures-door-beginnende-hardlopers/$file/Cijfersfactsheet%20Hardlopen.pdf

Wiehe, S.E., Hoch, S.C., Liu, G. C., Carroll, A. E., Wilson, J. S. & Fortenberry, J. D. (2008). Adolescent travel patterns: Pilot data indicating distance from home varies by time of day and day of week. *Journal of Adolescent Health* 42 (4), pp. 418-420. DOI: 10.1016/j.jadohealth.2007.09.018

Wong, D. (2009). The Modifiable Areal Unit Problem (MAUP). In Fotheringham, A.S. & Rogerson, P.A. (Eds.), *The SAGE Handbook of Spatial Analysis* (pp. 105-123). London: SAGE Publications Ltd.

Xin, Y. & S. Xiougang (2009). Linear Regression Analysis: Theory And Computing. Singapore: World Scientific Publishing Co. Pte. Ltd. First Edition.

Yoshimura, T. & H. Hasegawa (2003). Comparing the precision and accuracy of GPS positioning in forested areas. *Journal of Forest Research* 8 (3), pp. 147-152.

Appendices

Appendix 1: Data Permissions

Runner Tracks

Dataset content	GPS tracks of runners practicing for the Ladies Run or the Marathon of Eindhoven in 2015.
Data type	GeoJSON
Data features	Polyline features
Dataset owner	Technical University of Eindhoven & Fontys Sporthogeschool Eindhoven
Rights	By agreement of Prof. dr. Steven Vos (TU/e – Fontys) I can use this information for research purposes. I am not allowed to use or disclose this information for any other purpose.
Contract	1. This confidentially agreement is part of the research cooperation between TU/e, Fontys University of Applied Sciences and Utrecht University regarding the use of user-generated (big)data to analyse and monitor physical activity and sports participation in urban areas. This research is led by Prof. dr. Steven Vos. 2. In consideration of this collaboration and the opportunity for me to participate in this research and receive in connection therewith confidential information and data supplied by TU/e-Fontys, I, the undersigned, hereby agree that any and all Confidential Information and Data received by me shall be treated strictly as confidential and as proprietary to TU/e-Fontys (Prof. dr. Steven Vos). 3. I shall not, except as authorised in writing by Prof. dr. Steven Vos (TU/e – Fontys), a. disclose, reproduce, or use any Confidential Information and Data for other than the said purposes, or b. publish or authorise anyone else to disclose, reproduce, publish or use any Confidential Information and Data. 4. This restriction applies to all Confidential Information and Data which is not generally available to the public or to competitors. 5. Confidential Information and Data need not be kept confidential if: a. it is previously known to me other than by disclosure by TU/e – Fontys documented in writing, or b. it is or becomes generally known to the public (except if such public knowledge is the result of acts attributable to me), or c. Prof. dr. Steven Vos (TU/e – Fontys), explicitly agrees in writing that it need not be kept confidential. 6. At first written request, I shall promptly deliver to Prof. dr. Steven Vos (TU/e – Fontys), all drawings, manuals, letters, notes, notebooks, reports, and all other material of a private, proprietary or confidential nature containing Confidential Information and Data. 7. All Confidential Information and Data disclosed or submitted by TU/e – Fontys shall remain the property of TU/e – Fontys.

	The obligation for me to keep the Confidential Information and Data confidential shall expire three (3) years from the date it is received.

Fietsersbond Netwerk

Dataset content	Road network of the Netherlands of all roads a cyclists can go on, in addition to regional roads and highways. This dataset is enriched with additional information on the quality of the road and its surroundings.
Data type	esri Shapefile
Data features	Polyline features
Dataset owner	Fietsersbond Nederland
Rights	Permission to use this dataset for the purposes of this research, but nothing else.

Basic Information					
Name dataset	Fietsbondnetwerk_NL				
Publication date	2016				
Producer	Fietsersbond				
Data features	Polylines				
Projected coordinate system	RD_New				
Geographic coordinate system	GCS_Amersfoort				
Quality parameters					
Score	1	2	3	4	5
Completeness			X		
Compatibility					X
Consistency					X
Applicability				X	
Total score	17				
Standardized score	**0,85**				

Basic Information					
Name dataset	EHV_Clean				
Publication date	2015				
Producer	Technical University of Einhoven / Fontys Sporthogeschool Eindhoven				
Data features	Polylines				
Projected coordinate system	WGS1984				
Geographic coordinate system	WGS1984				
Quality parameters					
Score	1	2	3	4	5
Completeness					X
Compatibility			X		
Consistency				X	
Applicability					X
Total score	17				
Standardized score	**0,85**				

Basic Information					
Name dataset	Top10NL_WEGDEEL_HARTLIJN				
Publication date	2018				
Producer	Kadaster				
Data features	Polylines				
Projected coordinate system	RD_New				
Geographic coordinate system	GCS_Amersfoort				
Quality parameters					
Score	1	2	3	4	5
Completeness					X
Compatibility					X

Consistency					X
Applicability					X
Total score	20				
Standardized score	1,00				

Basic Information					
Name dataset	Criminality rates				
Publication date	2016 – 2017				
Producer	Statistics Netherlands				
Data features	CSV table				
Projected coordinate system	-				
Geographic coordinate system	-				
Quality parameters					
Score	1	2	3	4	5
Completeness				X	
Compatibility			X		
Consistency					X
Applicability					X
Total score	17				
Standardized score	0,85				

Basic Information					
Name dataset	Wijken en Buurten 2016				
Publication date	2016				
Producer	Statistics Netherlands				
Data features	Polygons				
Projected coordinate system	RD_New				
Geographic coordinate system	GCS_Amersfoort				
Quality parameters					
Score	1	2	3	4	5
Completeness					X
Compatibility					X
Consistency					X
Applicability					X
Total score	20				
Standardized score	1,00				

Basic Information					
Name dataset	Top10NL_TERREIN_VLAK				
Publication date	2018				
Producer	Kadaster				
Data features	Polygons				
Projected coordinate system	RD_New				
Geographic coordinate system	GCS_Amersfoort				
Quality parameters					
Score	1	2	3	4	5
Completeness					X
Compatibility					X

Consistency					X
Applicability					X
Total score	20				
Standardized score	1,00				

Basic Information					
Name dataset	Top10NL_WATERDEEL_VLAK				
Publication date	2018				
Producer	Kadaster				
Data features	Polygons				
Projected coordinate system	RD_New				
Geographic coordinate system	GCS_Amersfoort				
Quality parameters					
Score	1	2	3	4	5
Completeness					X
Compatibility					X
Consistency					X
Applicability					X
Total score	20				
Standardized score	1,00				

Basic Information					
Name dataset	Top10NL_WATERDEEL_LIJN				
Publication date	2018				
Producer	Kadaster				
Data features	Polylines				
Projected coordinate system	RD_New				
Geographic coordinate system	GCS_Amersfoort				
Quality parameters					
Score	1	2	3	4	5
Completeness					X
Compatibility					X
Consistency					X
Applicability					X
Total score	20				
Standardized score	1,00				

Basic Information					
Name dataset	Cden16_k8				
Publication date	2011 – 2016				
Producer	RIVM				
Data features	Asci Grid				
Projected coordinate system	-				
Geographic coordinate system	-				
Quality parameters					
Score	1	2	3	4	5
Completeness					X
Compatibility			X		

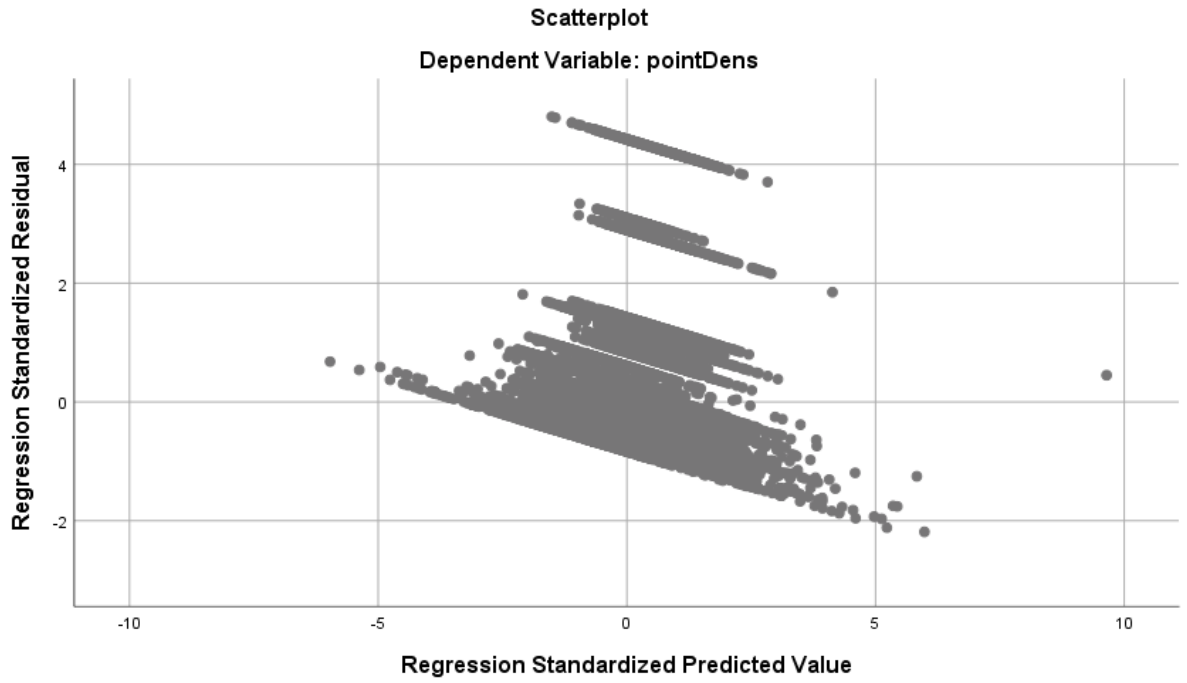
Consistency					X
Applicability					X
Total score	18				
Standardized score	0,90				

Basic Information					
Name dataset	PM10_luchtmeetnet, PM2,5_luchtmeetnet, NO2_luchtmeetnet, EC_luchtmeetnet				
Publication date	2016				
Producer	RIVM				
Data features	GeoTIF raster				
Projected coordinate system	-				
Geographic coordinate system	-				
Quality parameters					
Score	1	2	3	4	5
Completeness					X
Compatibility				X	
Consistency					X
Applicability					X
Total score	19				
Standardized score	0,95				

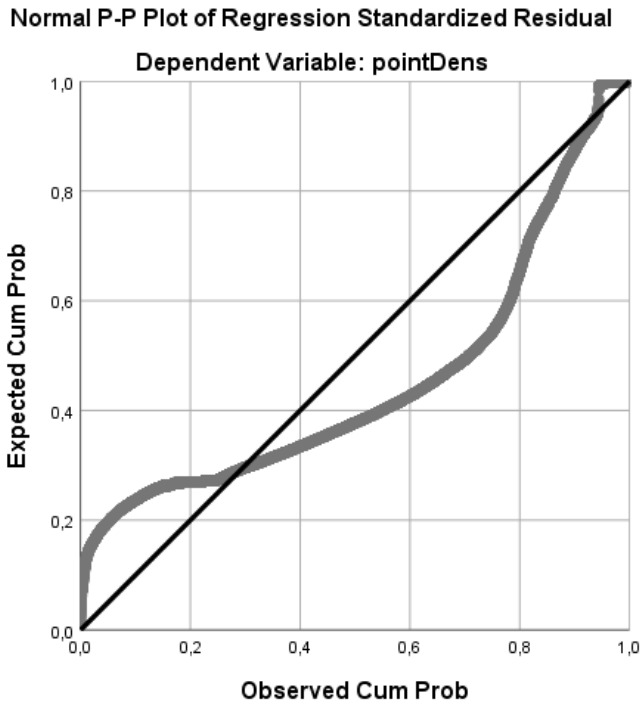
Basic Information					
Name dataset	Imergis_gemeentegrenzen_kustlijn				
Publication date	2016				
Producer	BrigGIS Geoservices BV				
Data features	Polygons				
Projected coordinate system	RD_New				
Geographic coordinate system	GCS_Amersfoort				
Quality parameters					
Score	1	2	3	4	5
Completeness					X
Compatibility					X
Consistency					X
Applicability					X
Total score	20				
Standardized score	1,00				

Appendix 3: Regression Analysis Assumption Testing Weighted Influences

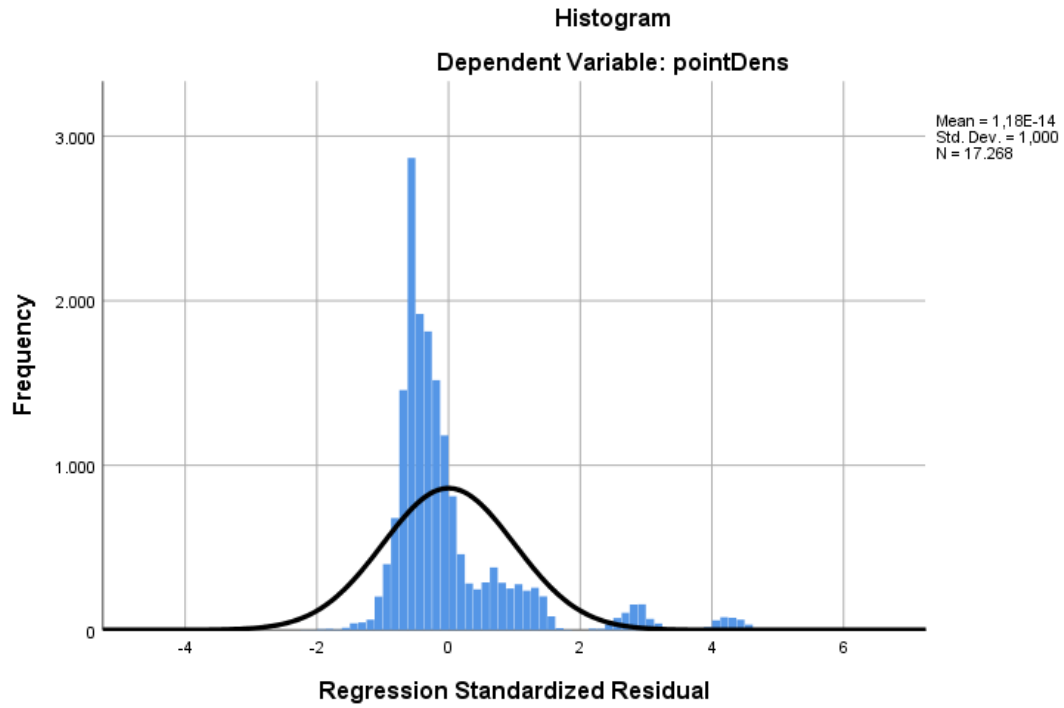
Homoscedasticity scatterplot



Normality of residues plot



Normality of residues histogram



Matrix of autocorrelation

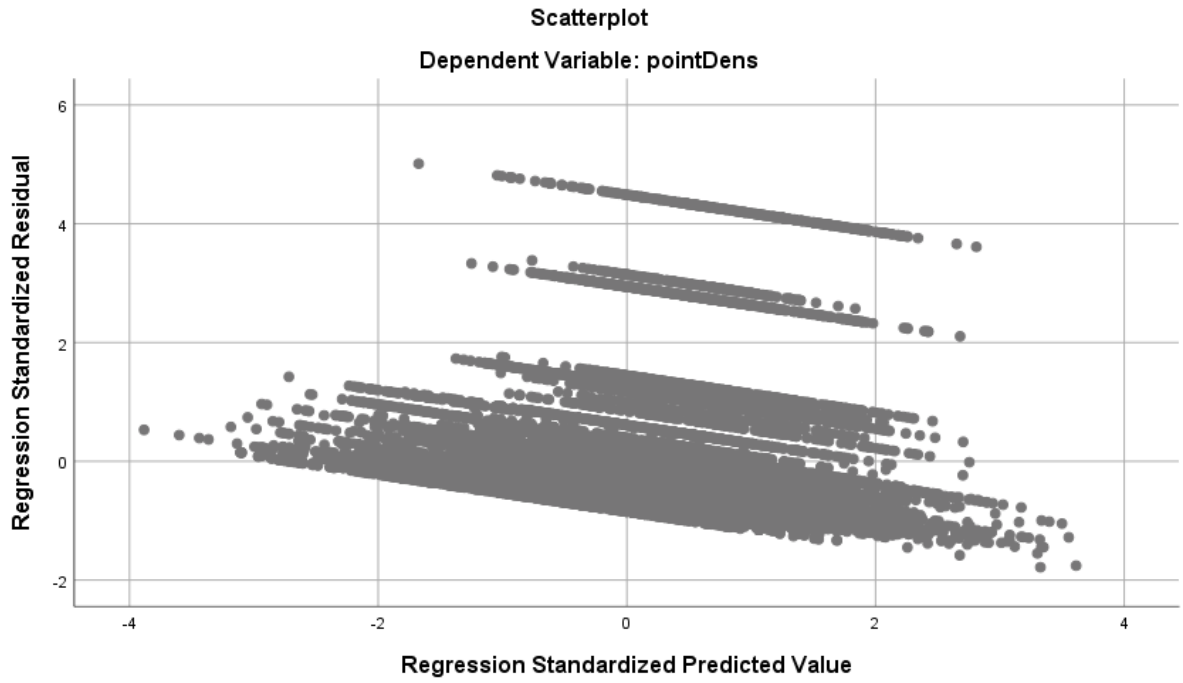
Coefficient Correlations[a]

Model			EV_infl	SQ_infl	NA_infl	LI_infl	MV_infl	VH_infl	CY_infl	SP_infl	AP_infl
1	Correlations	EV_infl	1,000	,033	-,100	-,017	-,018	-,043	,017	,334	-,272
		SQ_infl	,033	1,000	-,277	,094	-,047	,007	-,095	-,165	-,360
		NA_infl	-,100	-,277	1,000	,048	-,024	,203	-,048	-,098	,573
		LI_infl	-,017	,094	,048	1,000	,032	-,027	,078	,041	,012
		MV_infl	-,018	-,047	-,024	,032	1,000	-,118	-,166	-,053	-,177
		VH_infl	-,043	,007	,203	-,027	-,118	1,000	,306	,089	,282
		CY_infl	,017	-,095	-,048	,078	-,166	,306	1,000	,011	,030
		SP_infl	,334	-,165	-,098	,041	-,053	,089	,011	1,000	-,164
		AP_infl	-,272	-,360	,573	,012	-,177	,282	,030	-,164	1,000

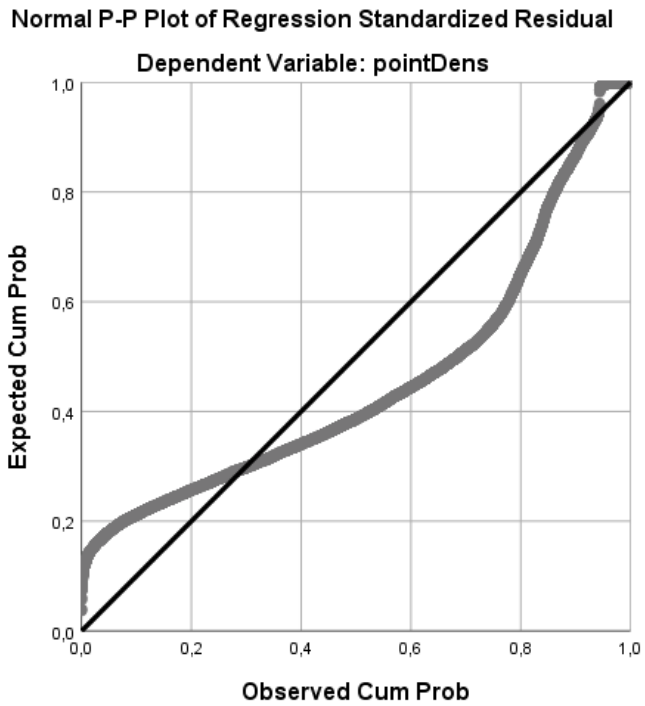
a. Dependent Variable: pointDens

Appendix 4: Regression Analysis Assumption Testing Weighted Influences

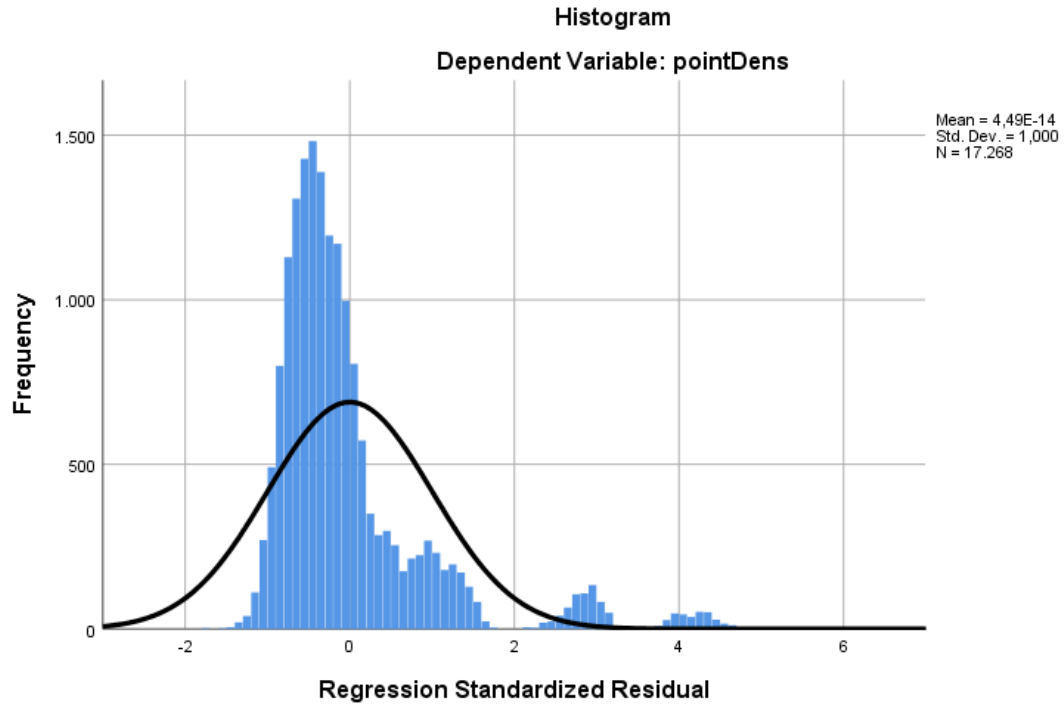
Homoscedasticity scatterplot



Normality of Residues Plot



Normality of residues histogram



Matrix of autocorrelation

Coefficient Correlations[a]

Model			EV_infl	MV_infl	VH_infl	LI_infl	SQ_infl	NA_infl	SP_infl	CY_infl	AP_infl
1	Correlations	EV_infl	1,000	-,034	,006	-,017	,043	-,119	,273	,022	,005
		MV_infl	-,034	1,000	,094	,024	-,035	-,029	-,103	-,180	,053
		VH_infl	,006	,094	1,000	-,002	-,027	-,144	,078	-,283	-,405
		LI_infl	-,017	,024	-,002	1,000	,123	,051	,017	,066	,060
		SQ_infl	,043	-,035	-,027	,123	1,000	-,286	-,142	-,113	,016
		NA_infl	-,119	-,029	-,144	,051	-,286	1,000	,004	-,059	-,019
		SP_infl	,273	-,103	,078	,017	-,142	,004	1,000	,010	-,465
		CY_infl	,022	-,180	-,283	,066	-,113	-,059	,010	1,000	-,010
		AP_infl	,005	,053	-,405	,060	,016	-,019	-,465	-,010	1,000

a. Dependent Variable: pointDens