Degrees of urbanisation in transport modelling

Analysing the impact of the spatial environment on travel behaviour using cluster analysis and propensity score matching

MSc Thesis

Janine Timmerman

ProRail

TU Delft

Degrees of urbanisation in transport modelling

Analysing the impact of the spatial environment on travel behaviour using cluster analysis and propensity score matching

by

Janine Timmerman

to obtain the degree of Master of Science in
Civil engineering with track Traffic and Transport Engineering
at the Delft University of Technology,
to be defended publicly on Monday August 19 at 16:00.

Student number:	4831578	
Project duration:	February, 2024 – August, 2024	
Thesis committee:	Prof. Dr. Ir. B. van Arem,	TU Delft, chair
	Dr. Ir. N. van Oort,	TU Delft, supervisor
	Dr. Ir. A. J. Pel,	TU Delft, supervisor
	Ir. J. G. Hogenberg,	ProRail, external supervisor

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Cover image from Pixabay (modified)

ProRail

TU Delft

Preface

Dear reader,

This report marks the end of my study at the TU Delft and the end of an intensive six months of working on this thesis. Although there are many more figures and analyses I would have liked to add, I am happy with the end result. And perhaps I have to admit that after writing almost a hundred pages of main text, it is okay to end it.

I could not have completed this thesis without the help I received. For that I want to express my thanks. First, I would like to thank ProRail for giving me the opportunity to do an internship there and my colleagues for making it a fun time. I hope my research provided valuable insights for you, even though the train was not the main focus. I want to especially thank Justin Hogenberg for being my supervisor at ProRail. After each meeting, you would encourage me by saying my latest results gave you confidence in the quality of the rest of the thesis.

Next, I would like to thank the remaining of my thesis committee at the TU Delft. Niels van Oort, the first time I contacted you is a little over a year ago, because I was looking for a topic for my thesis. I want to thank you for helping me through this process, bringing me into contact with ProRail, and giving me valuable feedback to improve my work. Adam Pel, I want to thank you for your useful meetings. I could always discuss my newest graphs and insights with and at end of the meeting I had new ideas of how to continue. Bart van Arem, I want to thank you for guiding me through this process. Your enthusiasm and encouraging words after the official meetings really helped me along.

Furthermore, I want to thank Remko Smit and Frank Hofman from Rijkswaterstaat and Jasper Willigers from Significance. You helped me with finalise my topic and improve my understanding of the LMS. I could always ask questions when there was something I did not understand.

I also want to thank my friends and family for being there for me and making my time as a student a fun and educational one. I especially want to thank my husband Antonie for supporting me and of course for helping with the spelling check (I apologize for the length of the thesis. I hope the snacks compensated a bit) and being my (sometimes involuntary) rubber duck when I did not know how to continue. At last, a small thanks to my cat Veggie for sitting many hours on my lap (or laptop) while I was trying to write code. Any weird spelling mistakes might be done by Veggie walking over my keyboahrsjkh;

Janine Timmerman
Delft, August 2024

Summary

Introduction and methods

Without accurate knowledge about future transport demand, it is difficult to make a good transport planning for the future (Profillidis & Botzoris, 2018, section 1.3). To make predictions for all of the Netherlands, the National Model System (Landelijk Model Systeem [LMS]) is used. This multimodal model makes predictions for the main road and rail network of the Netherlands and is an important tool for policy making. It is used by several organisations, such as the Dutch ministry of Infrastructure and Water management and ProRail (Ministerie van Infrastructuur en Waterstaat, 2023). One of the inputs of the LMS is data based on yearly travel surveys (Onderzoek Verplaatsingen in Nederland [OViN]).

The LMS is a complex model that is able to capture many different effects that can affect travel behaviour. However, no model is able to completely reconstruct reality and models are constantly being improved. The Netherlands is a small country, but still has many different regions, varying from very dense cities with a complex (public) transportation network, to rural communities where people are often dependent on their cars for mobility. All these differences in the spatial environment affect the way people travel (e.g. Kent et al., 2023; Cao et al., 2009), making it a challenge to capture the whole country with only one model. The LMS often uses the degree of urbanisation (DU) to differentiate between different regions. The DU is based on the population density. It is unknown how well the DU is able to separate the country in different regions, that each display different travel behaviour.

This research aims to gain insights in the differences in travel behaviour within regions with the same DU and between regions with a different DU. The goal is to investigate whether the differences between regions are sufficiently taken into account in transport models and, if needed, to give advice on how these differences between regions can be better implemented in those same models. To the author's knowledge, not much research exists about modelling travel behaviour between different regions with one large model. Sikder et al. (2013) recommends investigating whether it is possible to introduce a new variable that can replace the DU in the LMS, which could better distinguish differences in travel behaviour between different regions. This idea will be further researched in this thesis, by answering the following research question:

> *To what extent does the degree of urbanisation capture the difference in travel behaviour in different regions in current transport models and in what ways can these differences be captured more realistically with those same transport models?*

This thesis will primarily focus on the modal split and not on other aspects of travel behaviour like travel time and distance.

The first part of this thesis is a literature review and an analysis of the LMS documentation. This is done to find out which aspects of the spatial environment affect travel behaviour according to the literature and how the LMS has implemented these aspects in its model.

The second part of this thesis consists of a data analysis. First an exploratory data analysis is done to see what differences in travel behaviour between regions can be discovered and how they relate to the DU. Next, a cluster analysis is done to divide the zones in regions with similar spatial environment characteristics. This is done with the goal to find regions with similar travel behaviour that provide more insights than the DU. After that, a technique called propensity score matching is performed to discover to what extent differences in travel behaviour between clusters are caused by the spatial environment and to what extent they are caused by differences in demographic characteristics.

After these steps, an answer to the main research question can be formulated. See figure 1 for an overview of the methodology.

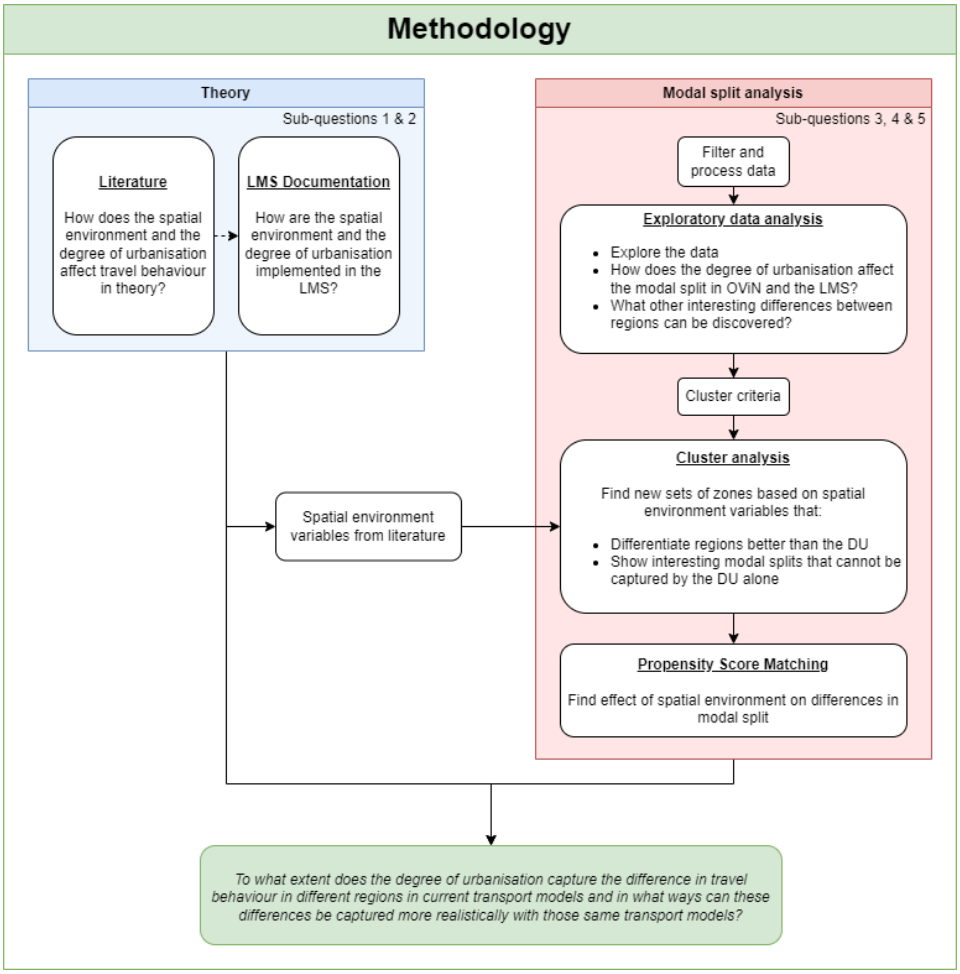

Figure 1: Schematic overview of the methodology to answer the main research question

Literature review and analysis LMS documentation

There is a lot of research about how the spatial environment affects travel behaviour. However, there is still debate on how much of these differences in travel behaviour between regions are caused by the spatial environment and how much is due to differences in demographic characteristics and preferences of people (Cao, 2014). A framework that is often used to quantify the spatial environment are the D-variables (Ewing and Cervero, 2010; Kent et al., 2023):

- Density: a variable per unit area, e.g. population density;
- Diversity: a measure for different land-uses in an area;
- Design: the characteristics of the street network;
- Destination accessibility: a measure of how easy it is to access different trip destinations;
- Distance to transit: the availability and quality of the public transport network;
- Demand management: measures that are meant to stimulate or dissuade the use of certain modes.

Demographics is sometimes added as a seventh D-variable (Ewing and Cervero, 2010). This variable is not part of the spatial environment, but it is important to control for it. The effect of the spatial environment usually gets less when demographic characteristics are taken into account (Cao et al., 2009). These D variables are only rough categories, can overlap and might be changed in the future. However, the D-variables are still a valuable tool to get a better overview of the different variables that

represent the spatial environment (Ewing & Cervero, 2010).

The core of the LMS consists of several modules. The most important modules were analysed to look for variables related to the spatial environment: the population module, car ownership module, travel frequency model and mode-destination-time of day-choice model. It was discovered that the D-variable framework is partly implemented in the LMS. Density and Distance to transit are implemented using many different variables. Especially the DU, which is a Density variable, is used many times in the LMS. Demand management is also implemented well. Variables that can be counted as Diversity and Destination accessibility are primarily related to jobs, instead of other aspects of the spatial environment, like land use diversity or accessibility to other points of interest. There are no explicit Design variables used, though they are presumably incorporated when determining the accessibility of origin-destinations pairs.

Based on these findings, it can be expected that the LMS will perform relatively well in modelling public transport modes in different regions. Besides that, modes of which the use depends a lot on population density are expected to be modelled more accurately.

Modal split analysis

First an exploratory data analysis is done to inspect the data and gain more insights in the differences in travel behaviour between regions and the ability of the LMS to model this. The most important findings are given below.

In general, the LMS seems to 'spread out' travel behaviour more than the data from OViN would suggest. In other words, it predicts similar levels of mode use in neighbouring zones. This can be a good thing (e.g. outliers are removed due to unreliable data), but it also removes different trends that are seen in the OViN data. The exploratory analysis suggests that on both national and regional level, the LMS overestimates the share of car driver trips and underestimates the share of walking trips. The average share of bike trips is modelled relatively well on national level when looking at the DUs and does not vary much. However, bike use can vary a lot when looking at zone level. By spreading out the travel behaviour, the LMS is unable to capture these differences accurately. Examples of this are found both in large cities and in rural areas, making this not only a problem for one type of region. Public transport use is modelled relatively well, although train travel is modelled more accurately than bus, tram and metro (BTM). Finally, places with the same DU do not necessarily display the same travel behaviour.

The second part of the modal split analysis is a cluster analysis. In the cluster analysis, sets of zones are grouped based on the characteristics of the spatial environment. The goal of this analysis is to find clusters that show more differences in modal split than clusters based solely on the different DUs and to identify regions with interesting travel behaviour, that cannot be captured by looking only at the DU. Zones are clustered based on the D-variables, as obtained from the literature. After that, the modal splits of the different clusters are analysed. At the end, two cluster sets are obtained. A weighted cluster set and an unweighted cluster set. Both cluster sets contain seven clusters. For the unweighted cluster set, as few variables as possible are used to make clusters, while still getting interesting results. The weighted cluster set is made with the assumption that each D-variable should have an equal weight but could be made up of several 'sub'-variables.

Both cluster sets are analysed and compared with the DU. The modal splits according to OViN and LMS were also compared. The most important findings are given below. In general, the LMS seems to predict the use of each mode better in the clusters where that mode is used the most, compared to the clusters were that mode is less popular. The modes car passenger, train and BTM are modelled the most accurately. The share of walking trips is often underestimated by the LMS, but the trends are modelled well (i.e. the share of walking trips according to the LMS increases and decreases in the same clusters as OViN). The share of car driver and bike trips seems to be captured the least accurately. The absolute and relative differences between the LMS and OViN are larger for the clusters, than for the DUs. This directly follows from the fact that the LMS has been calibrated to perform as well as possible with the variables that are implemented, which includes variables related to the DU.

The results show that some clusters have a similar DU, but a different modal split. The most interesting clusters that are found are the 'medium-sized city centres' and the 'suburbs of large urban

areas'. These clusters show interesting behaviour that does not follow the general trends that are seen based on the DU. The LMS seems to have the most trouble capturing the shares of car driver and bike trips for the medium-sized cities.

The last part of the analysis is propensity score matching. With this method, observations between two clusters with similar demographic characteristics are matched. This gives two new clusters that both have similar demographics. By comparing the differences in modal split before and after matching, the true effect of the spatial environment can be estimated. It was found that the spatial environment is on average responsible for more than 50% of the differences in modal split between regions. The effect of the spatial environment is the largest for BTM, walking and bike use. For these modes, it is presumably the most important to include enough spatial environment variables in the transport model. In general, the effect of the spatial environment is larger between the cluster pairs, compared to the different DUs. This indicates that the cluster sets are better than the DU at creating regions that show different travel behaviour based on differences in the spatial environment.

Discussion

The scope of the analysis was limited to the modal split, leaving out other aspects of travel behaviour (e.g. travel distance). Including those other aspects as well should result in a more complex analysis, but it would give a more complete picture. During the data analysis, several assumptions had to be made due to the lack of data. Besides that, several D-variables could have been determined more accurately in hindsight.

Because the size of the OViN dataset was not very large, especially when spreading the trips out over many zones, results that were obtained by analysing only a small set of zones might be less accurate. The cluster analysis was for a large part a manual process, because no way was found to optimize the process. This means that the cluster sets analysed in the thesis are not objectively the best cluster sets.

The results found in the modal split analysis are mostly in line with literature. Other studies found similar results using propensity score matching and general trends in the modal split that were observed during the analysis were often in line with results from literature.

However, there are also things that this study did different than existing literature. Similar studies often used 2-4 clusters to differentiate regions (e.g. Pot et al., 2023; Patnala et al., 2023), while this study argues that by using too few clusters, differences in travel patterns are lost. When using only 2 to 4 clusters, the analysis would focus mostly on rural versus urban areas. Other clusters like the suburbs and centres of large urban areas and medium-sized cities would never have been created. The results in this thesis showed that those clusters have a different modal split. This study is also different from many others, because it uses all the D-variables to evaluate the spatial environment, instead of only a few (Kent et al., 2023). Finally, many studies evaluate cycling and walking together, including studies in the Netherlands (e.g. Poorthuis and Zook, 2023; Van De Coevering and Schwanen, 2006). This study showed that there are significant differences between walking and cycling in different spatial environments.

The last part of the discussion is the generalisability. The methods used in this thesis can be be applied to other studies or can be used to evaluate other transport models. The exact results found in this thesis might not be directly applicable to other countries or other models.

However, the insights obtained from this thesis can help to better understand differences in travel behaviour between regions. It is important to include D-variables in transport models because differences between regions cannot be modelled by using only demographic characteristics. By using a cluster analysis combined with propensity score matching, specific regions can be identified with irregular travel behaviour that will need extra attention in modelling or research.

Conclusions and recommendations

This thesis aims to answer the following research question:

> *To what extent does the degree of urbanisation capture the difference in travel behaviour in different regions in current transport models and in what ways can these differences be captured more realistically with those same transport models?*

The DU is based solely on the population density. The data analysis showed that by looking only at the population density, important nuances in differences in travel behaviour are lost. The LMS can capture the trends in different regions for car passenger, train, BTM and walking well, although walking is underestimated. The different trends shown by bicycle and car use are often not captured well. Based on the results from the cluster analysis, the number of car trips that are modelled per cluster seems to be heavily affected by the DU, while bike use barely differs between different regions.

The differences in modal split between the different DUs and the different clusters are caused by both demographic characteristics and differences in the spatial environment. However, the effect of the spatial environment is larger.

The modal split modelling of the LMS could presumably be improved by adding additional D-variables to the model. To do this, it is important to keep in mind the following: In general, the share of car driver and bike will need the most improvement in capturing the right trends. The trends in walking are captured relatively well, though the absolute number of trips is underestimated. BTM and train are both captured relatively accurate. The biggest difference is that BTM use is affected more by the spatial environment than train use. Besides those points, the results showed that different cluster sets are able to uncover previously hidden trends in travel behaviour. These or other cluster sets could be implemented in the LMS in a similar way as the DU is currently used. It is likely that this will improve the ability of the LMS to distinguish between different regions, though this was not tested in this thesis.

The LMS is currently unable to accurately capture the effect of the differences in spatial environment. Policy makers and other users of the LMS should be aware that testing policies or future scenarios in the LMS that change aspects of the spatial environment might introduce additional uncertainties in the forecasts of some regions. As long as users of the LMS do not follow the forecasts blindly and critically evaluate the results, the LMS is still a very useful tool to get an idea of future transport. For ProRail the absolute number of train trips on an aggregated level is fairly accurate. On a smaller scale, the number of trips will become less accurate, but zones with high and low use are often correctly identified. These results can be kept in mind when using the LMS.

For future research, it would be interesting to look at more different aspects of travel behaviour and see if the obtained clusters still show significant differences. Similar research could also be done on a smaller region or a smaller transport model. Other elements that deserve further research is the effect of car ownership and population distributions on differences in travel behaviour; the effect of the spatial environment on different population segments or the effect of using different model structures to model different regions.

Contents

1 Introduction

1.1. Research context and problem

Many different transport models exist around the world. The goal of a transport model is to systematically and quantitatively analyse what happens when there are changes in a transport system. These could be changes due to external factors (e.g. the demography and land use change over the years) or internal changes in the transport system (e.g. change in timetable). The first transport models were developed in the 1950s, focusing on car traffic. Over time, these models evolved to larger and more complex models that focus on different kinds of modes and that are suitable for many different studies (Van Nes & De Jong, 2020).

Transport models give a prediction of future transport demand. This makes them crucial when designing new infrastructure, improving existing infrastructure and planning how the infrastructure is going to be used (e.g. how many trains have to drive on a certain section). Without accurate knowledge about future transport demand, it is difficult to make a good transport planning for the future (Profillidis & Botzoris, 2018, section 1.3). Some of these transport models only focus on a certain group of people in a small region, while other models try to forecast transport patterns for all people in a whole country.

For a countrywide model, a lot of data is needed in combination with a very complex model to capture the travel behaviour as realistically as possible. To make predictions for transport in the whole Netherlands, the Dutch national transport model (*Landelijk Model Systeem [LMS]*) is used. This model makes predictions for the main road and rail network in the Netherlands and is used in policy making of the Dutch Ministry of Infrastructure and Water management (*Ministerie van Infrastructuur en Waterstaat [IenW]*). Other uses of the model are to test the consequences of different choices in (infrastructure) projects. For example, what happens to the traffic intensities, travel times, noise and pollution when adding an extra lane to a certain highway. The results of the model can also serve as input for environmental analyses (Ministerie van Infrastructuur en Waterstaat, 2023).

ProRail is involved in the LMS by delivering detailed information about the timetables of the trains. For ProRail the predictions of the usage of the main rail network are very useful. These outputs combined with additional models, give information on the crowdedness on trains and the expected track occupation. ProRail can use these results to analyse where on the rail network bottlenecks might appear due to lack of capacity. (Hofman, 2017). ProRail expects an increase of 30 percent in passenger and freight transport in 2040. This increase should be handled properly and future bottlenecks should be prevented to keep the rail network accessible (ProRail, n.d.-a; ProRail, n.d.-b). Realistic predictions of the usage of the rail network and predictions of potential bottlenecks will help ProRail decide which sections of the rail network need extra improvements and investments and which sections will be able to handle the growing transport demand.

The LMS uses many different sources as input (Rijkswaterstaat, Water, Verkeer en Leefomgeving [RWS WVL], 2021b). One of these inputs is extensive survey data, which is complemented with additional data from registers and further processed to make the data more representative (*Onderzoek Verplaatsingen in Nederland [OViN]* and *Onderweg in Nederland [ODiN]*). For these surveys people all over the country are asked to fill in all the trips they made for one day, using a travel journal (Centraal Bureau voor de Statistiek [CBS], 2023).

The LMS is a complex model that is able to capture many different effects that can affect travel behaviour. However, no model is able to completely reconstruct reality and models are constantly being improved. The Netherlands is a small country but it still has many different regions, varying from very dense cities with a complex (public) transportation network, to rural communities where people are often dependent on their cars to get anywhere. All these differences in the spatial environment affect the way people travel (e.g. Kent et al., 2023; Cao et al., 2009), making it a challenge to capture the whole country with only one model. Cellissen et al. (2022) did an assessment of the most recent version of the LMS and found that car ownership in more urban regions is overestimated, while car ownership in more rural regions is underestimated. This is one example of differences between different type of regions that the model fails to completely capture. To take it one step further, large cities like Rotterdam and Amsterdam might seem similar (they are both dense urban regions), but Rotterdam is more car centered than Amsterdam which could potentially lead to different travel behaviour in these cities. Those differences in travel behaviour between cities should also be captured in the countrywide model. To give a small example, Cellissen et al. (2022) also looked at the growth of train stations. The growth of large stations in urban regions could be both overestimated (Den Haag Centraal) or underestimated (Utrecht Centraal). This is the result of a difference in travel behaviour between two highly urban regions that is not captured completely by the LMS.

The LMS uses utility functions to model travel behaviour (i.e. a person chooses the trip with the highest utility/ the trip that is the most beneficial for that person). These utility functions need a lot of input to make those predictions as accurate as possible. When differentiating between different regions, the degree of urbanisation (DU) is used in the LMS. (Regions with a low population density have a low DU and regions with a high density have a high DU.) The DU is not the only variable used to differentiate between regions (e.g. parking fee, job density), but it affects other variables, like the distribution of different household types in each region or the number of cars for each household in each region. This can make it hard to identify the effect of the DU or other zonal factors in the LMS (RWS WVL, 2021a; RWS WVL, 2021g).

It is unknown how well the DU is able to separate the country in different regions, that each display different travel behaviour. In other words, is travel behaviour in regions with the same DU similar enough that all the differences can be mainly explained by other variables, like the characteristics of the individual traveller? Likewise, is travel behaviour in regions with a different DU similar enough that existing differences in travel behaviour between two similar type of persons with the same travel motive can be mainly explained by the DU? Or are larger changes in the model needed to accurately capture these differences?

1.2. Objective and scope

This research aims to gain insights in the differences in travel behaviour within the same DU and differences travel behaviour for different DUs. This will be done with the goal to investigate whether the differences between regions are sufficiently taken into account in transport forecasting models and to give advice on how these potential differences in travel behaviour in different regions can be better implemented in those same models.

This research will primarily focus on analysing the OViN survey data from 2013 up to 2017 and the synthetic LMS matrices from the base year 2018. This means that the results of this thesis will be the most relevant for transport forecasting models that are similar to the LMS. Several studies have been done that analyse the accuracy of the LMS (e.g. Cellissen et al., 2022; Snelder and Vonk Noordegraaf, 2022; de Jong et al., 2008), but none of these studies specifically focus on the differences in travel behaviour within regions with a different or similar DU.

This thesis will focus on the modal split of the main mode, e.g. a person might use a train as main mode and uses a bike for access and egress. Only the train will be considered in this case. Other travel behaviour aspects that are determined in module 7.1 of the LMS will also be addressed to some extent. These are the travel frequency per day, the destination and the time of the day (RWS WVL, 2021g). Due to limitations in the scope, the latter aspects will be part of the literature review, but only appear briefly in the data analysis. Other aspects like route choice, driving behaviour or exact departure time will be excluded from the scope. Besides that, this thesis will analyse the observed (average) travel behaviour of the whole population for different regions (i.e. which travel patterns can be observed in

each region with certain characteristics) and will not analyse the decision making process and travel preferences of individual people.

This study will be done during an internship at ProRail. ProRail uses the results of the LMS when developing and planning the future rail network. However, they have less insights on the accuracy of the output of the LMS; what differences exist between different regions; and to what extent those differences are caused by the spatial environment and not by differences in demographics. ProRail knows that LMS predictions differ from reality when looking at the number of travellers per train station (Cellissen et al., 2022). However, they have less knowledge on how these differences in train trip predictions are related to other mode choices. With this research they hope to improve this knowledge.

1.3. Research gap and hypothesis

In a preliminary literature review, a research gap was identified and a hypothesis was formed of how transport models could be improved to better capture differences between different regions. The preliminary literature review can be found in appendix A. This section will provide an overview of the research gap and hypothesis.

To the author's knowledge, there are no studies that specifically focus on the ability of one large transport model to realistically model transport in a lot of smaller regions with different spatial characteristics, like the LMS does. Most research is focused on a smaller case study, like one metropolitan area. A possible explanation for this research gap could be that creating and analysing a model of that scope will require a lot time, money and data. These resources might not be available. Many studies, however, focus on spatial transferability. This is the ability of a model that was trained for one region, to model transport for another region (Sikder et al., 2013).

It is often unclear whether models are spatially transferable between regions or not. This could be due to the fact that the characteristics that separate those regions are unknown (McArthur et al., 2011). Mode choice and destination choice are especially difficult to transfer, due to differences in land-use, location preferences and mode availability between regions (Sikder et al., 2013; Linh et al., 2019).

Sikder et al. (2013) recommends to investigate whether it is possible to identify different region categories that separate different kind of regions from each other. These regions could include factors like land-use, demographics, features of the transportation network, etc. and could be an input variable in a transport model.

This idea will be further developed in this thesis, by researching whether it is possible to introduce a new variable that can replace the DU in the LMS and is better in distinguishing differences in travel behaviour between different regions. This variable should also be able to be used in other transport models besides the LMS, that currently mainly rely on variables related to the population density to distinguish between regions. The DU is currently used in many different places in the LMS. It is used as variables in the mode-destination-part of day discrete choice model, but it is also used on a higher level (e.g. when determining the population distribution for a zone). This will be further elaborated in section 3.3.

To conclude, little research exists about modelling travel behaviour between different regions with one large model. The lack of research in this area can be combined with the hypothesis of spatial transferability research: Creating region categories can help modelling travel behaviour in different kinds of regions. This forms the research gap and hypothesis for this thesis.

1.4. Research questions

The research objective of this thesis can be captured in the following main research question:

> *To what extent does the degree of urbanisation capture the difference in travel behaviour in different regions in current transport models and in what ways can these differences be captured more realistically with those same transport models?*

The first part of the main research question will focus on uncovering the effect of the DU on travel behaviour. The second part of the main research question has the implicit assumption that there are indeed differences in travel behaviour between regions that are not yet fully captured with the DU, and in extension, within transport models that primarily use this DU to differentiate between those regions. To help answer this main research question, several sub-questions were made.

- What region specific factors affect travel behaviour?
- In which ways are different travel behaviours in different regions captured in the Dutch national transport model?
- How does actual and predicted travel behaviour differ between regions with a different degree of urbanisation?
- How does actual and predicted travel behaviour differ between regions with a similar degree of urbanisation and what could be the cause of those potential differences?
- How can the Dutch national transport model be improved to capture the differences in travel behaviour in different regions more realistically?

The motivation for each sub-question and their corresponding research method will be further elaborated in the next section.

1.5. Research method

The first sub-question will be answered with the help of a literature review. This will be done with the goal to discover what is already known about travel behaviour and their relationship with the spatial environment. This information will provide valuable insights and will help to structure the data analysis that will be needed to answer the later sub-questions.

The second sub-question will focus on researching how the LMS has already included variables in their model related to the spatial environment. This will be done by studying the LMS documentation. This is needed to explain the different trends in travel behaviour that can be seen in the LMS and it will show how the regional factors included in the LMS can be compared with the current literature that was found for the first sub-question.

The third and fourth sub-question will require an extensive data analysis to answer. First an exploratory analysis will be done to see what differences in travel behaviour between regions can be seen and how they relate to the DU. For the fourth sub question, specifically, there can looked for examples of regions with a similar DU and see how their travel behaviour compares.

After the exploratory analysis, a cluster analysis will be done. This analysis will provide better insights for the third and fourth sub-questions, but also provide a basis for the fifth sub-question. The cluster analysis will attempt to divide zones in regions with similar regional characteristics that show similar travel behaviour, in line with the hypothesis of this thesis. After that, a technique called propensity score matching will be used to discover to what extent differences in travel behaviour are caused by the spatial environment and to what extent they are caused by differences in the demographic characteristics. The insights obtained from the data analysis will help formulating ways in which the LMS can be improved. This will give an answer to the fifth sub-question.

The first four sub-questions will provide an answer to the first part of the main research question and the fifth sub-question will provide an answer to the second part of the main research question. For this it is important to ensure that the insights obtained from this research are not only applicable to the LMS, but can also provide valuable insights for other transport models.

1.6. Outline report

This section will give the outline of this report. Chapter 2 will give the methodology of this thesis and will provide an introduction to the case study. The literature review and the analysis of the LMS documentation will be described in chapter 3. Chapter 4 will describe and give the results for the exploratory analysis, cluster analysis and propensity score matching. After that, chapter 5 will provide a discussion of the results and finally, the conclusions and recommendations are given in chapter 6.

2

Methodology

This chapter will present the research methodology for this thesis. First, a short introduction to transport modelling will be given to give some context and the case study (LMS and OViN/ODiN) will be introduced. After that, the methodology of the thesis will be described The first part of this thesis will consist of a literature review to find out how and to what extent the spatial environment affects travel behaviour and the second part will be a modal split analysis. The methodology for the data analysis in this chapter will not be in detail, because the specific details of this analysis are dependent on the outcomes of the literature review.

2.1. Introduction to transport modelling

As described in the introduction (section 1.1), the goal of a transport model is to systematically and quantitatively analyse what happens when there are changes in a transport system. Even though many different transport models exist, most of them have several elements in common. First of all, usually the study area of the model is split up in many zones. It is assumed that a trip begins in a zone and ends in a zone. There is detailed information about each zone, like the number of inhabitants and the number of jobs. These zones are connected with the help of a network. This could be a road network in the case of cars or bikes, or a public transportation network using lines and frequencies. In most transport models, the following four components can be discovered:

1. Trip generation - the number of trips leaving and entering each zone.
2. Trip destination - calculating for each origin zone the distributions of trips to all other zones (destinations). This gives an origin-destination (OD) matrix.
3. Modal split - the distribution of modes for each OD pair.
4. Assignment - the distribution of modes and trips over the whole network.

A final element is the trip purpose. Studies have found that incorporating trip purpose (e.g. commuting, leisure) in a model, gives a more accurate representation of reality (Van Nes & De Jong, 2020).

The type model described above is a trip-based model and is often referred to as a four-step model. Another type of transport model is an activity based model (ABM). It has some similarities to a more traditional four-step model (activities, destinations, modes and network assignment is determined), but it is based on the theory that people make (transport) decisions based on the activities they participate in. An ABM models a person's activities and travel choices across an entire day, considering the different kind of activities a person needs to participate in, and fitting it all in a schedule (Castiglione et al., 2014). An extension to the trip-based model is the tour-based model, which is a series of trips starting and ending at the same location. The LMS is a tour-based model, which will be elaborated in the next section.

2.2. Introduction case study

This section gives an introduction to the case study of this thesis: OViN survey data and the LMS.

2.2.1. General structure of the LMS

As described in the introduction, the LMS is a transport model that predicts the use of different modes (car, train, bus, metro/tram, (e-)bike and walking) and the usage of the main road and rail network in the whole of the Netherlands. The model is owned by Rijkswaterstaat (RWS), and the first version was developed in 1986. It is mainly used by the IenW, but it is also used by other ministries or other organizations like ProRail. The model is used in policy making, testing the consequences of different choices in (infrastructure) projects or serves as input for environmental analyses. ProRail mainly uses the output of the LMS, combined with additional models, to identify bottlenecks on the main railway network, which is important when planning the development of the railway network in the future (Ministerie van Infrastructuur en Waterstaat, 2023; Hofman, 2017).

The core of the LMS is the Groeimodel (GM, which can be translated to 'growth model'), which is responsible for making the synthetic OD matrices for each mode, motive and part of the day. The GM consists of seven modules. Each module is more or less self-contained, though they often use outputs from previous modules as input. The following four modules form the core of the GM (RWS WVL, 2021b):

- D4: The population module (in Dutch: Bevolkingsmodule): this module determines important population data for each zone, including car ownership.
- D5: Accessibility module (in Dutch: Bereikbaarheidsmodule): this module aims to determine the accessibility of the uncongested network.
- D6: Foreign traffic module (in Dutch: Buitenlandverkeermodule): this module determines the travellers that leave the Netherlands using the air or the ground.
- D7: Growth factor module (in Dutch: Groeifactormodule): this module calculates the growth factors to multiply the base matrices with to get an estimation of future transport matrices. The base matrices are detailed OD matrices based on real transport data and give an accurate estimation of transport flows for each mode in the base year. In the LMS version that is used for this thesis, the base year is 2018.

This thesis will focus mostly on module D7.1. This part determines the travel frequency, destination, mode and part of the day of for each type of person (RWS WVL, 2021g). This results in synthetic OD matrices for the base year and future years, that can be used to calculate the growth factors. For this thesis the synthetic matrix of the base year 2018 will be analysed and compared with 'real' travel behaviour according to OViN survey data.

In module D7.1, the four model components, as described by Van Nes and De Jong (2020), can also be found, though not exactly in that order. First, the travel frequency for each type of person and for each motive is determined (trip generation and trip purpose). Then, the destination and mode choice are determined roughly at the same time (trip destination and modal split) (RWS WVL, 2021g). The trip assignment to the network is determined in a later module and is not part of the scope of this thesis. It can be concluded that the LMS has many similarities to a more classic four-step model, but it is more advanced.

In the next subsection some definitions will be given of concepts that are used in the LMS. In section 3.3 a more elaborate overview is given of the LMS and the way variables related to the spatial environment are currently incorporated.

2.2.2. Definitions in the LMS

In this section a few definitions and concepts are defined that are used in the LMS and are relevant for this thesis.

Utility functions and nested logit

To determine different travel frequencies, modes, destinations and times of the day, the LMS uses utility functions in combination with a nested logit model (RWS WVL, 2021g). A logit model is a discrete choice model. Discrete choice models describe the choices of a decision maker among a set of alternatives. This set of alternatives is called a choice set and has three characteristics. The decision maker can only choose one alternative; the choice set should be exhaustive (each possible alternative must be included); and the choice set must be finite (Train, 2009).

For each alternative the utility is calculated, which is a measure for how beneficial the alternative is for the decision maker. The utility depends on characteristics of the alternative and characteristics of the decision maker. In the LMS, the decision maker is a type of person with a certain motive and the different modes and destinations are part of the choice set. The destinations are represented by non-overlapping zones that divide the whole research area, which makes the destination alternatives mutually exclusive, exhaustive and finite. In other words, the model determines the chance P_{HBmdtv} of person type t with origin H and motive m to choose destination B, during the part of day d, using mode v. This chance is determined for all possibilities. Before doing that, the travel frequency for each type of person t with motive m for origin H is determined.

The nested logit model is an expansion on the logit model and is able to capture correlations between different choices. The nested logit model calculates the probability for each alternative to be chosen by the decision maker (RWS WVL, 2021a). For more information about how (nested) logit models work, see Train (2009).

Tour

The LMS uses tours and not trips when modelling travel behaviour. A trip is defined as a movement from one location to another location. A tour is a series of trips that starts and ends at the same location. This way of modelling transport fits better with the theoretical idea that travelling is the result of people participating in activities at different locations, which is similar to the theory behind activity based models. There has to be a trip to the activity and a trip back home. It is possible to add secondary trips to a tour (e.g. a person visits a shop on the way from home to work. The main motive for travelling in this case is going to the activity: work). Besides a main motive, a tour also has a main mode (e.g. a person uses a train as the main mode, but also uses a bike for access and egress). There are a few advantages of using tours instead of trips. When someone uses a car to get to work, chances are big that the car will also be used on the way back. Secondly, a trip back happens always after the trip to an activity. This is difficult to model using a trip-based transport model. At a later stage, the LMS converts the tours back to trips (RWS WVL, 2021a).

Zone

The LMS has divided the Netherlands in 1406 zones, excluding the zones in neighbouring countries. The latter are outside the scope of this thesis. Each LMS zone is at least as large as the corresponding zone based on the four numbers of the Dutch postal codes (PC4), though often an LMS zone will contain several PC4 zones (RWS WVL, 2021a; RWS WVL, 2017). On average, a PC4 zone is 8.6 km^2 and an LMS zone 24.9 km^2. However, the sizes of an LMS zone can differ a lot depending on the location of the zone: from 0.12 km^2 up to 279 km^2. The LMS models trips between those zones, so the output of module D7.1 will be OD matrices that give the number of trips between each zone pair.

Degree of urbanisation

One of the ways the LMS distinguishes between different zones is the DU. The LMS has defined 6 different DUs, which are based on the population density. A degree of 1 corresponds with a very rural area, while a degree of 6 corresponds with a very highly urbanized area. Table 2.1 shows the corresponding population density for each DU.

The DU is determined in two steps. First, the population density of a zone is determined not only by taking the area and the total population of the zone itself, but also the area and population of the surrounding zones. Two zones are part of each other's surroundings if the euclidean distance between the centroids of the zones is equal to or less than 3 kilometres. The total area and the total population of a zone and its surrounding zones are used to determine the population density of a zone. The DU of a zone is then determined by using the surrounding zone (or the zone itself) with the highest population density. (RWS WVL, 2021a).

Table 2.1: The definition of the degrees of urbanisation according to the LMS (RWS WVL, 2021a).

Degree of urbanisation	Description	Population density [inhabitants/ha]
1	Rural area	≤2.5
2	Moderately rural area	≤6
3	Urban area	≤25
4	Very urban area	≤50
5	Large urban area	≤85
6	Centrum urban area	>85

2.2.3. OViN/ODiN

OViN/ODiN survey data is one of the inputs of the LMS. People all over the country from six years and older are asked to fill in some personal information and all the trips they made for one day using a travel journal. This data is then processed and supplemented with extra information from government registers. These surveys are conducted by the CBS on behalf of IenW. Such travel surveys have been conducted since 1978. From 2010-2017 this was done using OViN and since 2018 using ODiN. The goal of OViN/ODiN is to obtain information about the daily mobility of the Dutch population, which can be used for developing and testing transport policies (CBS, 2023). The version of the LMS used in this thesis with base year 2018 uses stacked OViN years from 2015 up to 2017 to calibrate its model parameters (RWS WVL, 2021a). The OViN/ODiN datasets contain data about the characteristics of the person who filled in the survey (e.g. age, gender, education), characteristics of its household (e.g. size of the household, household income, place of residence, number of cars) and information about all the trips they made on that day (e.g. mode, duration, distance, PC4 of origin and destination) (Centraal Bureau voor de Statistiek & Rijkswaterstaat [CBS & RWS], 2017c).

2.3. Overview of methodology steps

This section will give a global overview of the methodology of this thesis, see figure 2.1.

The methodology of this thesis consists of 2 parts. The ‘theory’ part will be a literature review and an analysis of the LMS documentation. This will provide an answer to the first two sub-questions. The second part is a modal split analysis, which will provide an answer to the last three sub-questions. The modal split analysis consists of an exploratory data analysis, a cluster analysis and propensity score matching. The results form the theory and the modal split analysis, together, form an answer to the main research question. Figure 2.1 gives a short summary of the goal of each step in the methodology. These steps and goals will be further elaborated in the next sections.

2.4. Theory

The first part of the thesis will focus on the theory. This part will answer the first two research questions and will provide all necessary information to set up the detailed methodology for the data analysis. The theory part of the thesis consists of a literature review and an analysis of the LMS documentation. Before the start of this thesis, a preliminary literature review was done to identify the research gap. This literature review, including its methodology, can be found in appendix A.

2.4.1. Literature review

The literature review, section 3.2, will cover the first sub-question: *“What region specific factors affect travel behaviour?”*. The goal of this section is to find a suitable theoretical framework that can be used to find and quantify factors related to the spatial environment that affect travel behaviour. In this section there will be a special focus on the DU and how this affects travel behaviour according to the literature. The result of this section will be a comprehensive overview of different (quantifiable) factors related to the spatial environment that affect travel behaviour.

For this part several search engines were used, mainly Scopus and ScienceDirect. The following keywords were inserted to obtain the intial set of papers:

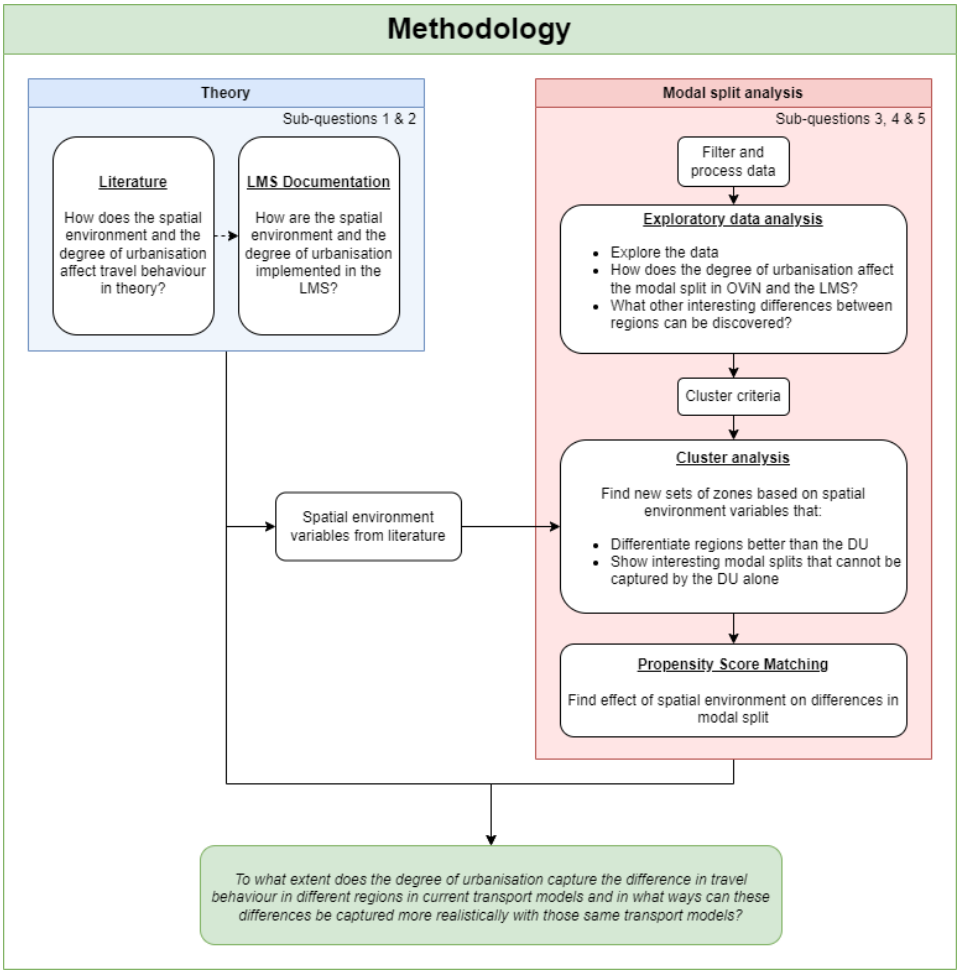

Figure 2.1: Schematic overview of the methodology to answer the main research question.

("degree of urbanisation" OR "degree of urbanization") AND "travel behaviour"

These keywords gave 5 hits in Scopus, 115 hits in ScienceDirect and 554 hits in GoogleScholar. These keywords were to broad for GoogleScholar and the results were discarded until better keywords could be formulated. To start, the abstracts and titles of the hits from Scopus and Sciencedirect were read through and based on this 3 papers from Scopus and 19 from ScienceDirect were chosen for further evaluation. This resulted in 4 relevant papers and 6 semi-relevant papers. However, with the help of snowballing (using the references from other papers) another 18 (semi) relevant papers were identified.

With this initial set of papers a theoretical framework was found, that is often used to quantify and analyse the effect of the spatial environment on travel behaviour. These were the 'D-variables', see section 3.2.2. This resulted in additional keywords to get a more comprehensive overview of the different factors affecting travel behaviour:

*"travel behaviour" AND "D-variable" AND [**specific d-variable(s)**]*

These keywords were inserted in Scopus and ScienceDirect. GoogleScholar was not used any further because it was assumed that Scopus, ScienceDirect and the snowballing technique already provided enough sources for an extensive literature review.

At the end around 40 relevant sources were used to do the literature review.

2.4.2. LMS documentation

The second part of the theory will focus on the second sub-question: *"In which ways are different travel behaviours in different regions captured in the Dutch national transport model?"*. This question will be

answered by studying the most recent documentation of the LMS and identifying all region specific factors that are used, including factors that only indirectly affect travel behaviour. These factors are then compared with the theoretical framework that follows from the first sub-question. This will show to which extent the spatial environment is currently included in the LMS and in which parts the LMS could be improved based on the literature. These findings can then be verified with the data analysis.

The documentation of the LMS was studied by reading through all available documents. Parts that were deemed less relevant were skimmed through. To make sure no relevant information was missed, several keywords were used to search through the (Dutch) documents using the search function: *'stedelijk-' (urban)*, *'zonal-'/'zone'*, *'urbanisatie-' (urbanisation)*.

2.5. Modal split analysis

This section will give a broad overview of the steps that were taken to perform the data analysis. Not all the exact choices that were made with regard to the data analysis and processing are handled in this chapter, because they require input from the literature review. Those choices will be elaborated in chapter 4. The goal of the data analysis is to provide the necessary information to form an answer to the three final sub-questions:

- How does actual and predicted travel behaviour differ between regions with a different degree of urbanisation?
- How does actual and predicted travel behaviour differ between regions with a similar degree of urbanisation and what could be the cause of those potential differences?
- How can the Dutch national transport model be improved to capture the differences in travel behaviour in different regions more realistically?

Because of limitations in the scope, the data analysis will primarily focus on the modal split. The reasons for this choice and some preliminary data analysis that was done on other aspects of travel behaviour can be found in the discussion (section 5.1).

Before the start of the analysis, the data must be filtered en processed to make sure the OViN data are in the same format as the LMS data. Based on the factors found in the literature, additional data must be gathered and processed. After the data processing some exploratory data analysis will be done. The objective of this analysis is to explore the data set, by showing initial statistics for the OViN and LMS data and to look for interesting differences or patterns that can be found. This step can be used to show the differences in travel behaviour between regions with a different DU (sub-question 3) and to search for areas that show interesting patterns in travel behaviour (e.g. areas with the same DU but a different modal split). These findings will help answering sub-question 4 and will provide extra guidance in structuring the cluster analysis, which is the second part of the modal split analysis.

The goal of the cluster analysis is to create regions based on characteristics of the spatial environment that are better in differentiating travel behaviour or regions that identify patterns that are not visible when looking only at the DU. This analysis will provide additional insights needed for the final two sub-questions and will show aspects where the LMS performs well and where is could still be improved.

After the cluster analysis, a technique called propensity score matching (PSM) will be used to quantify the effect of the spatial environment and the demographic characteristics on travel behaviour. The goal of this method is to discover if the spatial environment has a significant effect on differences in travel behaviour between different regions or if these differences are primarily caused by differences in the demography. This will help determining which parts of the LMS should be a focus when improving the LMS (e.g. adding additional region specific variables or improving the population distribution of each zone), providing additional information to answer the fifth sub-question.

All data processing and analysis was done using python. The python libraries that were used for handling the data and doing simple computations are *NumPy*, *Pandas* and *GeoPandas*. For the more difficult computations like clustering and regression, the *SciPy* and *scikit-learn* libraries were used. All the figures were made using the *Matplotlib* library.

2.5.1. Filtering and processing data

Before analysing the data, the data is processed and filtered.

Choice of years OViN/ODiN

The first choice is to select which years to use of OViN/ODiN. The LMS uses the years 2015-2017 to calibrate the model (RWS WVL, 2021a). Currently, OViN/ODiN datasets up to 2022 are available, but from 2020 to 2022 the travel patterns of the people are affected because of the travel restrictions due to Covid19. Because the LMS models travel behaviour for an 'average workday' with as base year 2018, datasets from 2020 onwards will not be used.

To increase the size of the dataset 5 years of data are together, which is 2 more years than used for the LMS. This is done to make the final dataset as large as possible, while not using too many years because travel behaviour in the Netherlands has changed slightly over time (Poorthuis & Zook, 2023). There are small differences in what data was gathered for OViN and ODiN, e.g. OViN selected people from all ages, while ODiN only uses people from six years and older. Additional data processing is needed to make OViN and ODiN consistent (CBS, 2018). To avoid this, only OViN data is used.

OViN contains data about the total trip of each person, but also data about the separate parts of the trip (e.g. an example of a trip is a person that moves from home to work with the train as primary mode. The separate parts of the trip include the person cycling to the station, sitting in the train and finally taking the bus to the office). To limit the scope, only the total trip and the main mode will be considered in the analysis.

The years that are selected for this thesis are the OViN datasets from 2013 up to 2017 (CBS & RWS, 2014; CBS & RWS, 2015; CBS & RWS, 2017a; CBS & RWS, 2017b; CBS & RWS, 2017c).

Stacking OViN years

After selecting which years to use, the OViN datasets need to be stacked. The following steps were done with the help of a memo that explains how to handle OViN data for workday analyses (RWS WVL, 2018a).

All trips that were made on a weekday or on a holiday were removed (including the Monday before or the Friday after a holiday if the holiday takes place on a Tuesday or Thursday). After that, all trips were removed that missed the PC4 for the origin and/or destination. Because this analysis will focus on the spatial environment, trips without any information about the locations cannot be used.

Each trip in OViN contains a 'FactorV'. This factor can be used to make the data representative for the population and get the total number of trips for the whole country. This thesis will focus on relative travel behaviour (e.g. the percentage of trips made by car) and not on absolute numbers. This makes the absolute value of FactorV unnecessary. However, this factor will still be used to calculate the relative travel behaviour in a region (e.g. trip A might be counted three times when determining the average travel duration, while trip B might only be counted once). The memo explains how to scale FactorV to a workday and to a base year, based on the number of workdays and the total population of that year. The latter is done to take population growth into account.

See equations 2.1, 2.2 and 2.3, where $FactorV_{wdx}$ is FactorV for an average workday in year x, $FactorV_{bjx}$ FactorV scaled to the base year, n is the number of workdays and $population$ the number of people living in the Netherlands in a certain year.

$$FactorV_{wdx} = \frac{FactorV_x}{n_x} \tag{2.1}$$

$$weightfactor_x = \frac{population_{baseyear}}{population_x} \tag{2.2}$$

$$FactorV_{bjx} = FactorV_{wdx} \times weightfactor_x \tag{2.3}$$

After that, the separate OViN years can be stacked to obtain one large dataset.

Match OViN with LMS zoning

In OViN the origin and destination locations of a trip are given on PC4 level. The LMS however uses its own zones, which are based on PC4 level, but are not exactly the same. To make the OViN trips comparable to the LMS data, each PC4 needs to be matched to an LMS zone.

RWS provided a data file, including documentation, to match each PC4 zone with an LMS zone (RWS WVL, 2017). However, in a few cases one PC4 zone was matched to two LMS zones or the PC4 zone was not included in the data file. A possible cause for the latter problem is that this postal code did not appear in OViN 2015-2017 which were used in the LMS, but only appeared in the years 2013-2014. To match the remaining PC4 zones a small algorithm was written that matched the geographic centre of the PC4 zone with the nearest LMS zone. When a PC4 zone was matched with 2 LMS zones, one of the LMS zones was chosen. If OViN trips are assigned to multiple zones, they would be counted double in the statistics, which is undesirable. The shapefile with data for the PC4 zones is from 2019 (Centraal Bureau voor de Statistiek & ESRI Nederland [CBS & ESRI Nederland], 2019). After matching all the trips to their corresponding LMS zone and removing the trips that could not be matched, the final version of the stacked OViN file contained **379,797** trips which were made by **115,396** individual persons. See section 4.3.1 and appendix D for more details about the matching process.

Find additional data for the LMS zones

Based on the findings of the literature review, additional zonal data is gathered about all the LMS zones, so the characteristics of the zones could be compared. The exact methods used and variables gathered will be described in section 4.1.2.

Most of the data gathered was open data from the CBS, which could be downloaded freely. However, most CBS data is available on PC4 level or neighbourhood level, which is even smaller than PC4 level. The neighbourhood zones were matched to LMS zones, in the same way as the PC4 zones were. The biggest difference was that it mattered less if a neighbourhood belonged to two LMS zones at the same time. (In the case of PC4, a trip could only be assigned to one LMS zone, otherwise it was counted double in the statistics.)

After each PC4 zone and neighbourhood zone was matched to an LMS zone, the zonal data had to be aggregated to get one statistic for each LMS zone. Some assumptions are needed for this. For example, the average household size needed to be determined for an LMS zone, while the data was available on PC4 level. The average household size was calculated using equation 2.4, where $HHsize$ is the average household size of a PC4 zone, HH the total number of households in that PC4 zone, x a specific LMS zone and i a PC4 zone belonging to LMS zone x.

$$HHsize_{LMS_x} = \frac{\sum_i HHsize_{PC4_i} \times HH_{PC4_i}}{\sum_i HH_{PC4_i}} \tag{2.4}$$

For this it was assumed that each LMS zone was made up fully of PC4 zones and that the whole PC4 zone was part of the LMS zone. In reality, this was not always true and a PC4 zone could also lie partly in another zone. For other statistics the weighted average was calculated in a similar way, though the ‘weight’ could also be the total population of a zone or the area and the PC4 zone could be replaced by the neighbourhood zone. In the case of neighbourhood zones, sometimes a neighbourhood was matched to two LMS zones. For the calculations, it was assumed that the neighbourhood was fully part of both zones.

Process LMS data

Processing the LMS data required less steps than processing the OViN data. OD matrices for each mode for the average working day were given. The only real processing that needed to be done was stacking the matrices for car driver and BTM (which were split for different day parts and/or motives) and removing all trips that started or ended outside of the Netherlands.

There has been made an attempt to also calculate the travel time and distances for each LMS trip, but that proved to be more difficult than originally thought, due to the lack of suitable data. This is further elaborated in section 5.1 and appendix G.

2.5.2. Exploratory data analysis

This first part of the data analysis, has not a very strict methodology and is mainly used to explore the data. The goal of this analysis is to answer sub-question 3, to make a start on sub-question 4 and to provide input for the cluster analysis. For the cluster analysis, criteria will be made which the clusters must satisfy. Insight obtained from the exploratory data analysis can serve as input.

First of all, this analysis looks at the differences in the modal split between the different DUs and between OViN and LMS at national level. After that, there will be zoomed in on several interesting areas to analyse the differences in modal split at zonal level. This will give more insights in how the LMS models the modal split in neighbouring zones with a similar or same DU and how this relates to the modal split as observed in OViN.

2.5.3. Cluster analysis and propensity score matching

After doing some exploratory data analysis, a more in depth analysis will be done using a cluster analysis. In the cluster analysis, sets of zones are clustered based on the characteristics of the spatial environment. The goal of this analysis is to find clusters that show more differences in modal split than the different DUs and to identify regions with interesting travel behaviour, that cannot be captured by looking only at the DU. Doing this analysis will provide valuable insights in the ability of the LMS to capture different trends in travel behaviour when zones are not clustered based on the DU. Besides that, this thesis aims to research whether it is possible to introduce a new variable that can be an improvement on the DU. These clusters can form the first attempt at creating such a variable.

Next, a way must be found to control for the demographic characteristics between the new clusters and the different DUs. This will be done using PSM. The goal of this method is to find out if the differences in modal split that can be observed between the clusters are (primarily) due to differences in the spatial environment or if most of those differences can be explained by differences in the demographic characteristics. These results can form an important justification for giving a final advice in how the LMS can be improved. For example, if most of the differences between regions are caused by the demographic characteristics, it might not be the best solution to add new regional variables in the mode choice logit model, because the population module of the LMS requires more improvements.

This section is structured in the following way. First, a method is chosen to perform the clustering. Next, the method to find the clusters is given and finally the PSM will be explained. Figure 2.2 shows a schematic overview of the methodology for the cluster analysis and PSM.

Hierarchical clustering

By using a clustering technique, zones with similar characteristics can be grouped together. By clustering based on the variables found in literature, travel behaviour according to OViN en LMS can be compared between the clusters.

This technique is in line with the one of the goals of this thesis: finding an alternative for the DU that is better in distinguishing differences in travel behaviour between different regions. If it is possible to find clusters that are an improvement over the DU, it would presumably be relatively easy to add those clusters to a transport model, in a similar way the DU is used at this moment. Because the DU also divides the LMS zones in clusters, it will be insightful to compare the clusters made based on D-variables and the clusters based on the DU.

A disadvantage of using clustering is that the individual effect of the d-variables on travel behaviour is less clear. To mitigate this, the distribution of the d-variables within the clusters will be compared and analysed.

A method called ‘hierarchical clustering’ will be used to cluster the zones, following several other studies (e.g. J. Liu et al., 2024; Park et al., 2018). The first step of hierarchical clustering is to define a dissimilarity measure between each observation pair. A common used measure is the euclidean distance. At the start, each observation is treated as its own cluster. The hierarchical clustering algorithm will iterate over these observations and each time, merge the two observations that are the most similar until there is one cluster left. This gives an upside down tree shaped graph, called a dendrogram. This dendrogram is ‘cut’ at a certain height to get the desired number of clusters (James et al., 2023, pp. 503–556). This thesis will use the euclidean distance as dissimilarity measure and use the ‘Ward’s’ method to determine the dissimilarity between two groups of observations, which is also called: ‘linkage’. For more information about how hierarchical clustering works, see James et al. (2023, pp. 503–556).

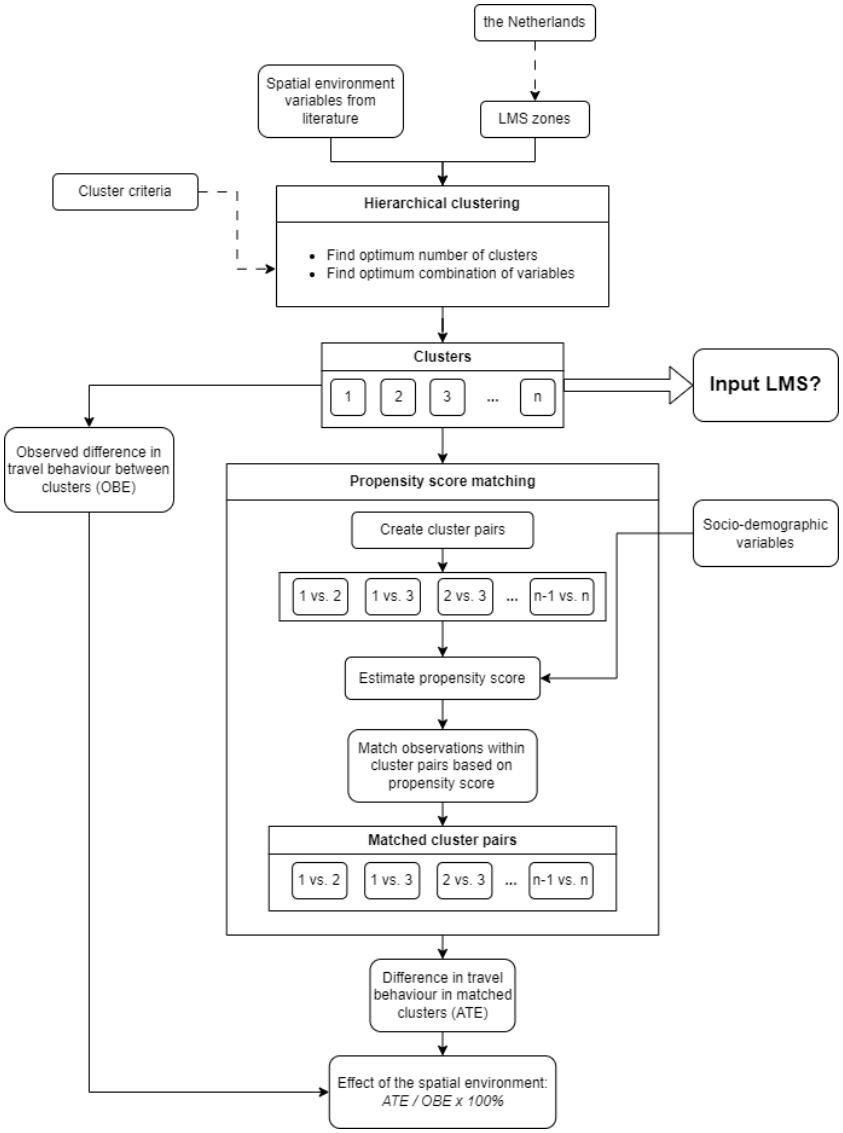

Figure 2.2: Schematic overview of the methodology for the cluster analysis.

Method to find the best clusters

Hierarchical clustering needs 2 different inputs: the number of clusters and the variables that need to be clustered. Both need to be determined.

For the number of clusters, there are several indicators that can help finding the best cluster. However, there is often not one clear number of clusters that is the best. An indicator that is often used in similar studies is the silhouette score (e.g. J. Liu et al., 2024; Pot et al., 2023). This value measures the similarity of an observation to observations in its own cluster, compared to the observations in other clusters. The number of clusters with the highest silhouette score is considered to be the best (Rousseeuw, 1987). Other examples of indicators that can be used are the Calinski-Harabasz score, which is also maximized for the optimal number of clusters (Calinski & Harabasz, 1974), or the Davies-Bouldin score, which has to be minimized (Davies & Bouldin, 1979).

However, using these indices blindly does not necessarily give the best number of clusters for the goal of this thesis. The optimal number of clusters should not be too low (e.g. if only 2 clusters are defined: ‘urban’ and ‘rural’, it is very unlikely that these clusters are an improvement on the 6 DUs), but also not too high (e.g. defining 20 different clusters will likely result in overfitting due to the limited size of the data. This will not be useful to implement in transport models).

The final number of clusters will be decided based on what the author believes captures the different

regions in the best way without underfitting or overfitting, using the different indices as additional input.

Next, it must be determined which spatial environment variables to include in the clustering. It is important to limit the number of variables, to make it easier to implement the clusters in a transport model (e.g. when making clusters based on 100 variables. the clusters might be able to capture the different regions better, but it can be a very time-consuming process to gather all the data for maybe only a small improvement in the clusters. It would also be more difficult to figure out the effect and usefulness of specific variables).

There is not one clear indicator which can be maximized (or minimized) to get the optimal solution. The clusters will be made and evaluated manually by the author to decide which variables should be included. However, there are several factors that will help in this process:

- The variables can be sorted based on their correlation with travel behaviour. This can help finding variables that are likely to improve the clusters.
- The correlation between the variables can be determined. This can help deciding which variables to include (e.g. if two variables are highly correlated it might not be needed to include both).
- After a cluster is made, the distribution of the spatial environment variables within the clusters can be evaluated to see how it affects the clusters (e.g. if a variable has similar values in the different clusters or has a very high variance, it might not be useful).
- During the exploratory data analysis, regions can be detected that have very similar or very different travel behaviour. To qualify the clusters, it can be checked if these regions fall into the same clusters or not.
- Look at the variance in travel behaviour within the clusters (i.e. zones within a cluster should have similar travel behaviour) and the variance of the means of travel behaviour between the clusters (i.e. the average travel behaviour between the clusters should differ as much as possible).
- A cluster should not contain too few zones to prevent overfitting. There will not be a specific limit on the cluster size (e.g. if there is a cluster for city centres that cluster would probably be small). However, it is important to be extra critical of small clusters and judge their relevance.

Other criteria for the clusters can be decided after the exploratory data analysis.

Propensity score matching

The different clusters that are made using the hierarchical clustering, presumably have different demographic characteristics. To make sure the differences in travel behaviour found between different clusters are caused by the spatial environment, those demographic characteristics have to be controlled for. This should also be done with the different DUs. This way the DUs and the clusters can be compared.

In similar studies this is commonly done with propensity score matching (PSM) (e.g. J. Liu et al., 2024; Park et al., 2018; Pot et al., 2023; Mokhtarian and Van Herick, 2016; Cao et al., 2010; Patnala et al., 2023). The general idea of PSM is to select similar observations of two clusters, in this case observations that have similar demographic characteristics. Observations that cannot be matched to any observation of another cluster are removed. This way, randomization is mimicked and at the end there are two clusters that have similar demographic characteristics and the differences in travel behaviour can be compared (Cao et al., 2009). A disadvantage of this method is that it is only possible to compare two clusters at once. So when the number of clusters increases, the number of analyses needed, increases even faster. One of the clusters is assigned to the ‘treatment’ group and the other to the ‘control’ group (Patnala et al., 2023).

The methodology of PSM consist of three parts. First, the propensity scores must be estimated. This is the probability of an observation belonging to the treatment group, based on the demographic characteristics (J. Liu et al., 2024). This can be estimated, for example, using a binary discrete choice model (e.g. logit) (J. Liu et al., 2024) or logistic regression (Pot et al., 2023).

Secondly, observations in both clusters are matched using the propensity score. For this a caliper length is used, e.g. a caliper of 0.01 means that the maximum difference in propensity scores of two

matched observations can be 0.01. After that, both clusters have similar demographics. It is assumed that there now is controlled for residential self-selection (Patnala et al., 2023).

It is important to check if the demographics between the clusters are similar after performing the PSM. This can be done using the standard mean difference (SMD), see equation 2.5.

$$SMD = \frac{100(\overline{X}_T - \overline{X}_C)}{\sqrt{\frac{S_T^2 + S_C^2}{2}}} \tag{2.5}$$

Here, $\overline{X}_T$ and $\overline{X}_C$ represent the mean values of a demographic characteristic for the treatment and control group, and S_T and S_C represent their standard deviations (J. Liu et al., 2024). A value of $|SMD| < 10\%$ indicates that the matched clusters are similar enough (Oakes & Johnson, 2006).

Finally, the impact of the effect of the spatial environment can be calculated. The average treatment effect (ATE) is the impact of the spatial environment on travel behaviour and can be obtained by calculating the difference in the mean values of the travel behaviour indicators between the clusters (e.g. difference in car use) after PSM. The observed effects (OBE) are the differences in the mean values of the travel behaviour indicators between the clusters before PSM. The ratio of ATE to OBE represents the true effects of the spatial environment (J. Liu et al., 2024). This is, of course, under the assumption that the demographic characteristics that were used in this analysis are extensive enough to capture the differences in population and attitudes. To check if the different ATE and OBE values are statistically significant, an independent t-test is used.

It is possible to perform PSM on all different clusters and use ATE or the ratio of ATE to OBE as a measure when searching for the optimum combinations of variables. However, due to the computational power needed to perform a single PSM, this was not feasible and it was only used to evaluate the final clusters and the DUs.

2.6. Improving LMS

Based on the results of the exploratory data analysis and the cluster analysis, advice can be given on how the LMS, or other similar transport models, can be improved.

First of all the clusters itself: if it is possible to create clusters that are better in capturing differences in travel behaviour than the the clusters made by the DU, it can already provide a relatively easy way to improve the LMS.

Secondly, based on the insights obtained from the literature, the study of the LMS documentation and the whole data analysis, more general advice can be given on how the LMS or similar models can be improved to capture the differences in regions better.

3

Literature review and analysis of the LMS documentation

This chapter will present the literature review that is conducted for this thesis. First, a definition is given for the degree of urbanisation (DU). The second section focuses on travel behaviour and the interaction between travel behaviour and the spatial environment. There will be some attention to factors that affect travel behaviour unrelated to the spatial environment. This gives more context and aims to give a better understanding of travel behaviour in general. The goal of this section is to provide the information to answer the first sub-question and to provide input for the data analysis. The third section gives an analysis of how aspects of the spatial environment are incorporated in the LMS and how they presumably affect model outcomes. This is done using the LMS documentation. Finally a summary is given and the LMS documentation is compared with the insights form the literature review.

3.1. Degree of urbanisation

The DU is a measure of how 'urban' a region is. There are many different ways to define the DU and there is not one standard worldwide way to define it, though it is often based on the population density (Taubenböck et al., 2022). The Statistics Netherlands (CBS, n.d.) uses five different DUs, based on the density of addresses, where a DU of 1 corresponds with a highly urbanised area and a DU of 5 with a rural area. The LMS, however, uses six different DUs based on the population density, where a DU of 1 corresponds with a rural area and 6 with a highly urbanized area (RWS WVL, 2021a). Because there is not one clear definition for the DU, for this thesis when talking about the DU, the definition according to the LMS will be used, unless specified otherwise.

Many studies identify the DU as a factor affecting travel behaviour (e.g. Susilo and Maat, 2007; Poorthuis and Zook, 2023). A lot of other studies do not use this exact term, but have identified the population density as a factor (e.g. Thao and Ohnmacht, 2020; Van De Coevering and Schwanen, 2006).

The DU can affect travel behaviour in several ways. First of all, regions with a different DU might have different travel alternatives (e.g. trams and metros are usually only found in very dense areas), a different network quality (e.g. dense areas might have less roads that are suitable for cars and more roads meant for cyclists and pedestrians) or a different kind of population (e.g. people might move to denser areas, because they prefer travelling by public transport). This thesis focuses on aspects that are related to the spatial environment and not on aspects related to the different demographic characteristics. However, these demographic characteristics are an important factor and should not be forgotten. Residential self-selection is the effect that individuals start living in locations that align with the way they prefer to travel (Ettema & Nieuwenhuis, 2017). This will be further elaborated in section 3.2.3.

3.2. Travel behaviour

The goal of this section is to find out what kind of factors affect travel behaviour and how travel behaviour is related to the spatial environment. This section focuses on different aspects of travel behaviour to give a more complete overview, and not only at the modal split.

First a theoretical overview is given about the kind of factors that affect travel behaviour. After that, literature about the effect of the spatial environment will be reviewed and finally an overview is given on the extent of the effect of the spatial environment and how it relates with the different demographic characteristics.

3.2.1. What is travel behaviour?

This section gives an overview of the kind of factors that affect travel behaviour. Not all of this is part of the scope, but it will help placing this thesis in a larger context.

According to Van Acker et al. (2010) travel behaviour can be seen as the result of several hierarchical decisions. Short-term travel decisions are based on the daily activities a person undertakes. Those activities are often on separate locations (e.g. work, shopping centre, visiting friends), so there is a need for travel. Sometimes, a person cannot participate in the activities they want, because they do not have the opportunity. This can lead to medium-term travel decisions like moving or changing jobs. Finally, there is lifestyle which is a long-term decision. A lifestyle is based on, among others, persons beliefs, attitude and motivations. Medium- and short-term travel decisions are made based on long-term travel decisions. An example of such a long-term decision is having kids, which might lead to moving to a child-friendly neighborhood and doing activities like bringing your child to school. The description of travel behaviour above is based on the perspective of an individual. It is important, however, to consider the context around the individual. A person will have family, friends, colleagues (the social environment) and will live in a neighborhood, which is also part of a larger spatial context (the spatial environment).

Short-term travel behaviour choices consist of several decisions. A person has to decide in which and in how many activities at different locations they participate. After that they will need to decide where they will perform the activity (the destination choice) and which mode they will use to travel to the activity, and back. They will need to decide at which time to depart and the route they will take (Bovy et al., 2006). Those choices are the results of both reasoned and unreasoned influences. When someone makes a trip for the first time, the reasoned influences probably play a larger role, while repeated behaviour (habits) depend more on unreasoned behaviour i.e. the behaviour becomes automatic (Van Acker et al., 2010). In transport models, it is often assumed that the traveller (the decision maker) makes rational choices and tries to optimize his personal situation (utility maximization) (Bovy et al., 2006).

This thesis focuses on short-term travel decisions and the effect of the spatial environment on these decisions. However, it is important to keep the larger context in mind when looking at the spatial environment. It is not possible to explain all travel behaviour by looking only at the short-term decisions and the spatial environment. Even though there is a lot of disagreement to what extent the spatial environment affects travel behaviour, most studies agree that it affects travel behaviour to a certain extent (Cao et al., 2009). Besides the spatial environment and demographic characteristics, factors like the economic context, policies and the culture also play an important role in understanding why people display certain travel behaviour (Kent et al., 2023).

3.2.2. Spatial environment

There is a lot of research about how the spatial environment affects travel behaviour and, by extension, also about how the spatial environment can be changed to affect travel behaviour (e.g. make people use other modes than car). Several of the characteristics that are assumed to affect travel behaviour are called the D-variables: Density, Diversity, Design, Destination accessibility and Distance to transit. Sometimes Demand management is included and Demographics. This last variable is not part of the spatial environment, but it is important to control for it. Those D variables are only rough categories, can overlap and might be changed in the future. However, they can still help to get a better overview of the all the possible variables (Ewing & Cervero, 2010). The relationship between the different characteristics of the spatial environment and travel behaviour is not always straightforward. Large cities might be

denser and have more diversity in land-use than rural areas. These different characteristics are related and interwoven, and can have a combined impact on travel behaviour (Dieleman et al., 2002). This means it might not always be possible to separate the effects of the different characteristics of the spatial environment. Even though the D-variables are widely accepted and used, a lot of studies include only a few of the D-variables in their research. Kent et al. (2023) did a literature review of 104 studies. 30% included more than 3 different D-variables and only 4 studies included all 6 D-variables (not counting Demographics). This shows that there is still room for more research that includes all of these D-variables.

It is not necessarily true that spatial environment characteristics that displayed a certain effect in one country, also have the same effect in another country. The magnitude of certain factors can be very different in different countries and sometimes the same factors can be responsible for reverse effects in different countries (Van De Coevering & Schwanen, 2006). This can make it useful to include more research in the literature review that was done in the Netherlands. This does not mean that research about other countries is excluded. It still plays an important role in better understanding travel behaviour.

The remaining of this subsection gives an overview of the effect of the different D-variables. An overview of the different D-variables that were found, is given in table 3.1.

Density and the degree of urbanisation

Density is defined as a variable per area unit. Possible choices for the variable are population, jobs, number of certain buildings or the building area (Ewing & Cervero, 2010). The DU is also a measure of Density, and to be more specific: often a measure of the population density (Taubenböck et al., 2022). See section 3.1 for a more elaborate definition of the DU. There are several reasons why the (population) density can have an effect on travel behaviour. With a higher population density more activities can be reached with active modes like walking and cycling; more different kind of services can exist in a smaller area, reducing travel distances; the distance between jobs, homes and other activity locations can be reduced; and a more extensive public transportation network is more profitable (Stead & Marshall, 2001).

In the Netherlands the average travel distance decreases with a higher DU (Schwanen et al., 2002; Susilo and Maat, 2007), while the average travel time (slightly) increases (Van Der Hoorn, 1979; Schwanen et al., 2002; Susilo and Maat, 2007; Poorthuis and Zook, 2023)[1]. This implies that the average speed decreases with a higher DU. These trends are seen both in the daily averages and the averages per trip; across different motives; and across different years. A possible explanation for a lower speed for cars is the increased congestion in more urban areas (Schwanen et al., 2002).

In the UK the average travel distance both by car and for all modes decreases as the population density increases (Dargay and Hanly, 2003; Stead and Marshall, 2001). Thao and Ohnmacht (2020) found that in Swiss, the average daily travel distance by car decreased when the population density increased and also when the job density increased. The population density has a stronger effect on car distance than the job density. Van De Coevering and Schwanen (2006) did research based on data from several cities in the US, Canada and western Europe. They also found that the average travel distance by car decreased when the population density increased.

This does not mean that the same patterns could be found in each region. When looking at the US, Canada and Europe separately, the population density in the US seems to have very little effect on the travel distance by car (Van De Coevering & Schwanen, 2006). In Nova Scotia Millward and Spinney (2011) found that not only the travel distances in the inner city were lower than in more rural areas, but also the travel time. This is in contrast with the studies in the Netherlands. These same results can be found in Nanjing, China (Feng et al., 2013).

There is less evidence on the effect of the spatial environment on the travel frequency. Existing literature often found only a limited or no relationship between frequency and Density (Van Der Hoorn,

[1] It is important to note that not all sources mentioned define the DU strictly based on the population density. Schwanen et al. (2002) bases its 'degrees' on population density, land-use mix and structure of the urban system, which results in 4 different 'degrees' for regions in the Randstad and 2 outside the Randstad. Van Der Hoorn (1979) also does not give a clear definition of the DU, but gives 7 different degrees with only a label as explanation (e.g. rural areas, commuter towns, medium-sized cities). Susilo and Maat (2007) and Poorthuis and Zook (2023) both use 5 DUs based on address density, similar to the definition of the CBS.

1979; Stead and Marshall, 2001; Thao and Ohnmacht, 2020). Dargay and Hanly (2003) found no significant relationship between density and frequency in the UK, except for the densest areas. There, the frequency was a bit lower than average. Millward and Spinney (2011), however, found in Nova Scotia that the trip frequency in the densest areas was a bit higher than average.

Differences in mode choice for different densities has been researched more. For the Netherlands research finds that, in general, the share of cars decreases with a higher DU/ population density, while the share of public transport increases. (Schwanen et al., 2002). These trends also hold true when looking at only a subset of all trips, like long distance trips (Limtanakool et al., 2006)[2] or family visit trips (Rubin et al., 2014)[3]. Poorthuis and Zook (2023) shows that the share of car use in urban areas has been decreasing between 2004 and 2020, while it remains steady in non-urban areas[4]. The share of walking and biking is also higher in urban areas and has been increasing over time, while it slowly decreases in non-urban areas. According to Schwanen et al. (2002) potential car drivers in more urbanised areas use cycling, walking and public transport instead of the car.

Different results can be found when looking only at the subgroup of Dutch children (4-11 years). They walk and use public transport more with higher DUs, though they cycle more in less urbanised areas. Furthermore, no relation could be found between the DU and travelling as a passenger in a car (Kemperman & Timmermans, 2012)[5]. This means that the exact same trends (decrease in car use in urban areas) can not be found in each subset of trips, even when looking only at the Netherlands.

In the UK there is an increase in trips by public transport and walking and a decrease in car trips in areas with a higher density (Stead & Marshall, 2001). Van De Coevering and Schwanen (2006) researched travel behaviour in several cities in Europe, the US and Canada and also found that walking increased, while car use decreased with the population density. For public transport they found a positive relationship with public transport use and job density. The effect of Density on mode choice in other countries seems similar to the Netherlands.

In studies in the Netherlands, the DU is often based on the population density. Puylaert et al. (2022) uses instead of the DU, as defined by the CBS, a so-called 'proximity index' (in Dutch: Nabijheidsindex) which is based on both the population density and the job density. When comparing mode choice frequencies across different DUs and across different levels of the proximity index, a larger variety of travel behaviour was seen. This indicates that the proximity index might be better at capturing differences in travel behaviour then the DU.

Ewing and Cervero (2010) did a meta-analysis of existing travel behaviour literature of mostly the US, with the goal to quantify the effect sizes of the different spatial environment characteristics. They did not find a strong relationship between Density and travel behaviour. They argue that this could mean that Density is more of an intermediate variable and that it could be expressed using the other D variables like Diversity or Design.

Diversity

Diversity is a measure for the different land-uses in an area. A possible way to measure Diversity is with entropy measures that give a high value for areas with a lot of variety in land-use and a low value for areas with areas with less variety (Kockelman, 1997). It is also possible to use measures like job-population ratio, though they are less popular (Ewing & Cervero, 2010).

Limtanakool et al. (2006) uses several land use indices in their study in long distance trip in the Netherlands, including the entropy measure. (Other studies (e.g. Harts et al., 1999; Kockelman, 1997) also defined several indices. A few of them will be used later in this thesis. For more details about the indices and the equations, see section 4.1.2.) They found that the entropy measure is positively correlated with using the train for commuting. In other words, commuters are more likely to take the train to work when their workplace is close to many other types of facilities. Similar results were found for long-distance business and leisure trips. Feng et al. (2013) found that in the Randstad the entropy measure is negatively correlated with travel time. In other words, when the land-use mix is higher, the

[2] Limtanakool et al. (2006) uses 4 levels based on population density.
[3] Rubin et al. (2014) uses 3 DUs based on address density.
[4] Poorthuis and Zook (2023) counts the two highest DUs, according to the CBS definition, as urban areas. This consists of about 50% of the population. The remaining areas are counted as non-urban.
[5] Kemperman and Timmermans (2012) uses 5 DUs based on address density (similar to the CBS definition).

travel time and travel distance are shorter. When the land-use mix is higher, there are more opportunities within a small distance, which could decrease both travel time and distance. The effect of land-use on travel distance is larger than on travel time.

Van De Coevering and Schwanen (2006) introduced a less common measure of Diversity. They argue that the historical development of an area are important to travel behaviour. To include this in their research, they use the proportion of houses that were built in certain time frames. They find that the land use characteristics of the city centre, that was built before 1945, are more closely related to travel patterns than the population density of the whole city.

Næss et al. (2017) studied travel distances in two urban areas in Norway. To account for Diversity, they included the job to workers ratio. They found a slight reduction in travel distance for local residents, when there was a local job surplus. However, a lot of employees are non-local and it does not reduce their driving distance. The literature review by Stead and Marshall (2001) stated that the studies that were done that looked at this job to workers ratio found only a small effect on travel behaviour. The job to workers ratio had a small effect on mode change and commuter time, and a slightly larger effect on travel distance. Ewing and Cervero (2010) found in their meta-analysis that the job to workers ratio has a larger impact on walking than the entropy measure, suggesting that people are more likely to walk by bringing jobs closer to homes, than by making the land-use more mixed.

Design

The Design variable says something about the characteristics of the street network. There are a lot of different ways to measure this. It is possible to look at characteristics like the number of intersections; the proportion of sidewalks or other type of roads; the widths of the roads; if there are a lot of straight connected roads, or more curving indirect roads; etc (Ewing & Cervero, 2010).

For example, Ewing and Hamidi (2015) looked at the intersection density and the percentage of four-way intersections in the US. Both variables were negatively correlated with the total travelled distance.

Sung and Eom (2024) uses road network density (the length of road per area), road ratio (percentage of land used for roads) and the average road width to quantify Design. They found that railway access increased with a higher road network density. They also theorized that a higher road density could reduce distances for pedestrians, making walking a more viable option. Li et al. (2024) also uses road network density and found indeed that by increasing the road density in Chinese urban villages, the share of walking and cycling was increased, while the total distance travelled was decreased.

L. Liu et al. (2023) found that proper streetscape planning (e.g. height-width ratio of the streets; is the ‘street wall’ continue or are there large gaps between buildings) does not only increase the visual appeal of an area, but it can also reduce car travel and increase the chance that other modes are used.

Destination accessibility

The Destination accessibility variable measures how easy it is to access different trip destinations. A possible way to measure it is to use the distance to the city centre or the number of destinations that can be reached within a certain travel time (Ewing & Cervero, 2010).

Næss et al. (2017) found that in some urban areas in Norway the Destination accessibility has the largest effect on car travel distance, compared to the other D-variables. They measured this variable by looking at the distance between home and the city centre. People who live farther from the centre travel longer distances on average. This was also found in Copenhagen, especially when looking at the distance of commuter trips (Næss, 2006). They conclude that living closer to the city centre, helps reducing the amount of daily travel. Similar results were found in the US and China. Ewing and Hamidi (2015) found that in the US the distance to the city centre and the accessibility of jobs by car had significant effect on the travel distance by car. T. Liu and Ding (2024) showed that in Beijing the distance to the city centre was one of the most important factors affecting travel distance by car, together with the distance to the metro station and the household income.

Thao and Ohnmacht (2020) looked at the distance from home to several services (e.g. schools, cinemas) and places of nature (e.g. forests, parks) and found that the daily distance travelled by car is lower when those services and nature locations are closer.

Distance to transit

The Distance to transit variable is often measured as the average of the shortest routes from home or work locations to the nearest train station or bus stop. It is also possible to look at the number of train stations or bus stops in an area or the distance between stops (Ewing & Cervero, 2010). Kent et al. (2023) argues that it is important to not only look at the availability of a transit stop, but also at the quality of the public transport. For example, the frequency, reliability and the accessibility to the rest of the network.

Thao and Ohnmacht (2020) found in Swiss that a higher quality public transport system[6] increases the frequency of walking and cycling trips and decreases the frequency of car trips. The frequency of public transport trips increases with the the quality of the public transport system.

Besides only the quality, the built year of especially rail based transit is also important. Sung and Eom (2024) found that if the rail transit service was built after houses were already occupied, it was more difficult to alter travel patterns because people already had preferred modes.

Demand management

Travel Demand management are measures that are meant to stimulate or dissuade the use of certain modes. Examples of Demand management are the restriction of car parking, or restrictions of cars in general. It could also include monetary measures like parking fares (Kent et al., 2023). Another example is the commuter allowance employers pay their employees. Often, travel Demand management is meant to provide an incentive to use more sustainable modes of travel like public transport instead of private cars (Sung & Eom, 2024).

3.2.3. How large is the effect of the spatial environment?

While many studies have researched the different aspects of the spatial environment that affect travel behaviour, there is still a lot of debate how much of these differences in travel behaviour are caused by the spatial environment and how much is due to demographic characteristics and the preferences of people (Cao, 2014). Ettema and Nieuwenhuis (2017) studied residential self-selection in the Netherlands and found that the spatial environment has an independent effect on travel behaviour and residential self-selection only plays a limited role. According to Hong et al. (2013) the effect of the spatial environment differs based on the exact location and also based on travel motive. Based on a literature review, Cao et al. (2009) found that the spatial environment does have an effect on travel behaviour, even when taking self-selection into account. However, they say the effect of the spatial environment, when there is controlled for residential self-selection. So without including self-selection the effect of the spatial environment might be overestimated. The few studies that quantified the effect of the spatial environment and that of residential self-selection, found most of the times that the effect of the spatial environment is stronger than that of residential self-selection.

Næss (2006) questions the conclusions of studies that only found weak relationships between the spatial environment and travel behaviour. They argue that weak relationships that were found might be because the assumptions that were made in the model did not manage to capture the real influence of the spatial environment. Another reason could be that the studies did not include the correct variables. For example, some studies compared two areas with different Density and Design variables, but did not include the location of the areas compared to the centre of the city or region.

Furthermore, Van Acker et al. (2011) researched mode choice for leisure trips and found that ignoring attitudes could even lead to underestimating the effect of the spatial environment.

Cervero (2002) argues that it is very important to include land-use variables in models for mode choice in urban areas, even if a variable only serves as a proxy. Otherwise it might be possible that the effect of another variable is overestimated. For example, someone is researching mode choice and is making a utility function for bus choice. They include a variable for transit travel time, but do not include a variable for population density. The travel time in the bus is presumably correlated with population density, because the distances are shorter and the waiting time is also shorter due to a higher frequency, so the population density could only be a proxy for the travel time. However, there are also other variables that are closely correlated with population density (e.g. bus stops in high density areas

[6] Thao and Ohnmacht (2020) bases the quality of the public transport system on the distance to the service and on the quality. It distinguishes 5 different levels of public transportation quality.

might provide shelters that protect against weather). The effect of these unknown variables related to population density are not included in the utility function, but can still have an effect on mode choice. So by excluding the variable for population density, the effect of transit travel time might be overestimated because it also includes the effect of other variables correlated with population density.

To conclude, both the spatial environment and residential self-selection have an effect on travel behaviour, though the strengths of these effects are still being discussed. The most important conclusions for this thesis are that the effect of residential self-selection should be taken into account in the data analysis and in answering the main research question. Besides that, it is important to include enough different variables for the spatial environment, even though they might serve as a proxy.

3.2.4. Conclusions

To conclude, there are many different factors related to the spatial environment that affect travel behaviour. This literature review identified the theoretical framework called the D-variables to help structure and quantify the spatial environment: Density, Diversity, Design, Destination accessibility, Distance to transit, Demand management and Demographics. An overview of the D-variable and how these variables appear in literature and the LMS can be found in section 3.4.

3.3. Region specific factors in the LMS

This section will focus on the LMS and explain which factors related to the spatial environment are used in the LMS. These factors will be connected to the D-variables as identified in the literature review. This will give a better idea to what extent the spatial environment has been taken into account. First, a description is given of the the modules D4 (population module) and D5 (accessibility module). These modules prepare important data that are needed for the central part of the GM, module D7 (growth factor module), to work. This module determines the travel frequencies and the mode, destination and time of day choice using utility functions and a nested logit model.

For a more general overview of the LMS, GM and its modules and some definitions, see section 2.2. In this section only the relevant modules, that include the spatial environment, will be studied. This means that module 6 (foreign traffic module) will be excluded. This thesis will only analyse trips that start and end in the Netherlands.

3.3.1. Population (module D4.1)

This subsection is based on the LMS documentation for module D4.1 (‘Programma QUAD’), RWS WVL (2021c). When other sources are used, it will be explicitly mentioned.

Because the GM works on the level of individual persons and households, it is important to get an accurate overview of the the households and the population in each zone. This is done by defining different household types, that are based on several characteristics (e.g. household size, the number and type of workers, age and income). In total, there are 378 different household types (RWS WVL, 2021a). However, based on existing data in the Netherlands, it is not possible to know the household distribution for each zone exactly for the base year or the future years. There are, however, several other sources:

- OViN data: besides information about trips, it also contains much data about the exact compositions of households.
- 28 variables of each zone, that gives information about totals. For example: the number of persons in each age group, the number of part time workers for each gender or the number of households. These values are called targets.

These data can be used to make the a-priori household distribution. This is an average distribution of each household type for each DU (RWS WVL, 2021a). It can also be used to determine the average of the targets of each zone for each household type. This will give a better idea of what each household type looks like.

The goal of this module is to generate a distribution of household types for each zone. This distribution should be as similar as possible to the targets of the zone, but also deviate as little as possible from the a-priori household distribution. This makes it a minimisation problem, which is separately solved for each zone.

This also implies that the household distribution in each DU is similar to some extent. However, it is possible to use weights in the optimisation problem to make some targets (and in extent the a-priori household distribution) more important. The total number of households is the most important target. Cellissen et al. (2022) suggests that future versions of the LMS should make the a-priori household distribution more important in determining the household distribution of each zone. Due to the importance of the targets, the current results are not always realistic. They found that during the minimisation, some household categories are 'emptied', while others are disproportionately increased to meet the target values. They also advise to make an a-priori household distribution for each zone for the base year (instead of for each DU that is done now).

To conclude, module D4.1 bases its household distributions mostly on demographic data of each zone, with the exception of the DU. However, based on the findings of Cellissen et al. (2022) the effect of the DU on the household distributions is low. When looking at the D-variables, only the Density variable is represented in this module to a small extent (and of course Demographics).

3.3.2. Car ownership (module D4.2)

Car ownership is an important predictor for travel behaviour (Dieleman et al., 2002). The LMS uses car ownership data from a model called DYNAMO The LMS receives a few car ownership values and coefficients for a model. The values are determined outside the GM. The goal of module 4.2 is to update these coefficients and make them consistent with them GM. The coefficients will later be used in Module 7 to model car ownership. It also gives as output some information about car ownership for each zone. For this module, many variables are used, including variables related to the spatial environment. These will described in this section. The section is based on the documentation for module D4.2, RWS WVL (2021d). When other sources are used, it will be explicitly mentioned.

Variables related to the spatial environment

As said above, only a few values are known about (future) car ownership: The total number of households with one car, with two cars, with more than two cars and the total number of cars. These 4 values are for the whole Netherlands. Besides that, there are initial car ownership coefficients for all variables. These values that are initially only known for the Netherlands as a whole, will be predicted for each zone separately, including the number of households without a car.

Most of the variables that are used are related to the demography of a zone (e.g. age of the head of the household; number of persons without a license). The variables related to the spatial environment are the following:

- There are two variables for disposable income based on DU. One variable is disposable income for a household that lives in an area with one of the three lowest DUs, while the second variable is the disposable income for a household that lives in an area with one of the three highest DUs.
- Population density of all zones within a radius of 1 km of the gravity point of the concerned zone. (The DU is based on the population density with a radius of 3 km.)
- There are two variables for the job density, one using zones in a radius of 1 km and the other in a radius of 5 km around the gravity point of the zone.
- Parking fee per hour.
- The maximum number of parking permits per household (if lower than 4).
- A dummy variable for a zone without parking limitations (4 or more parking permits per household).
- Two dummy variables for the DU. One for household with a DU of 3 or 4, and one for a household with a DU of 1, 2 or 3.

- The ratio of jobs in agriculture and the total number of jobs.

In practice, the initial car coefficient are the same as the output, except for the ASC (alternative specific constant). Those are updated in this module. These constants can be used to capture effects that are not included in the list of variables and are specific to a certain choice.

Effect of the spatial environment

Looking at the initial values of the car coefficients, an estimate can be made about their effect on car ownership. These initial values can be found in table 2.1 of the documentation (RWS WVL, 2021a). The relevant parts of the table can be seen in appendix B

In total, there are 4 different alternatives. A household with 0 cars, 1 car, 2 cars or 3 or more cars. The coefficients show that car ownership is positively correlated with income (a higher income increases all alternatives with 1 or more cars) and that the correlation is stronger when the household has a lower DU. The population density is negatively correlated with car ownership (a higher population density decreases the chance of ownership of 1 or more cars) and also negatively correlated with job density. The dummy variables for the DU show that rural areas increase the chance of having more cars and living in an area with a DU of 3 or 4, decreases the chance of having 0 cars, but says nothing significant about the number of cars.

The ratio of jobs in agriculture has a negative correlation with having 0 cars. The parking fee is positively correlated with having 0 or 1 car and the number of parking permits (including unlimited) is positively correlated with having 1 or more cars.

To conclude, most variables related to the spatial environment that affect car ownership are related to Density (DU, population density and job density). There are however also some variables related to Diversity (ratio of jobs in agriculture) and Demand management (parking fee/ permit).

The signs of the coefficients are not unexpected. According to the literature review car ownership is positively related with car use, and car use negatively correlated with density (Van Acker et al., 2014). The coefficients for car ownership show that car ownership is also negatively correlated with density. A parking fee (an incentive to discourage car driving) is positively correlated with having less cars and having no parking restrictions is positively correlated with having one or multiple cars. Locations with more agriculture, are presumably in more rural areas, in which car use is also higher in general, according to the literature.

It is important to keep the results of these car ownership coefficients in mind when looking later at the frequency, mode, destination and time choice models. It might seem at first glance that car ownership is a variable not related to the spatial environment, but at closer inspection the models does use several D-variables in determining car ownership.

3.3.3. Accessibility (module D5)

The goal of the accessibility module is to determine the quality of the accessibility for each OD pair, including intra zonal trips. This accessibility can be expressed in values like cost, travel time and distance (RWS WVL, 2021e).

The exact way these accessibility values are determined will not be covered in this report. However, there are presumably D-variables (implicitly) used for this. Especially the Design variable will be related to values like actual travel distance, compared to the euclidean distance (e.g. a higher road density presumably means more direct routes).

3.3.4. Introduction Sample Enumeration System (Module D7.1)

Module 7.1, Sample Enumeration System (SES) is the most important part of the LMS. It determines the transport demand. First determines the travel frequency for each type of person and motive and then the mode, destinations and the part of day will be determined. This is done using a nested logit model with many different variables. The variables related to the spatial environment will be covered in his section. The section is based on the documentation for module D7.1, RWS WVL (2021g). When other sources are used, it will be explicitly mentioned.

The following steps are done for each type of person and each zone:

- The attraction for each mode, destination and part of day alternative is determined, i.e. the utility for each alternative.
- The utility for all alternatives together can be used to calculate the accessibility of the origin zone. This is called the 'logsum'.
- Based on several variables (including the logsum), the number of trips originating in a zone is calculated.
- Based on the utilities of each alternative, compared to the other alternatives, the trips are divided over different destinations, modes and parts of day.

An overview of the structure of SES can be seen in figure 3.1.

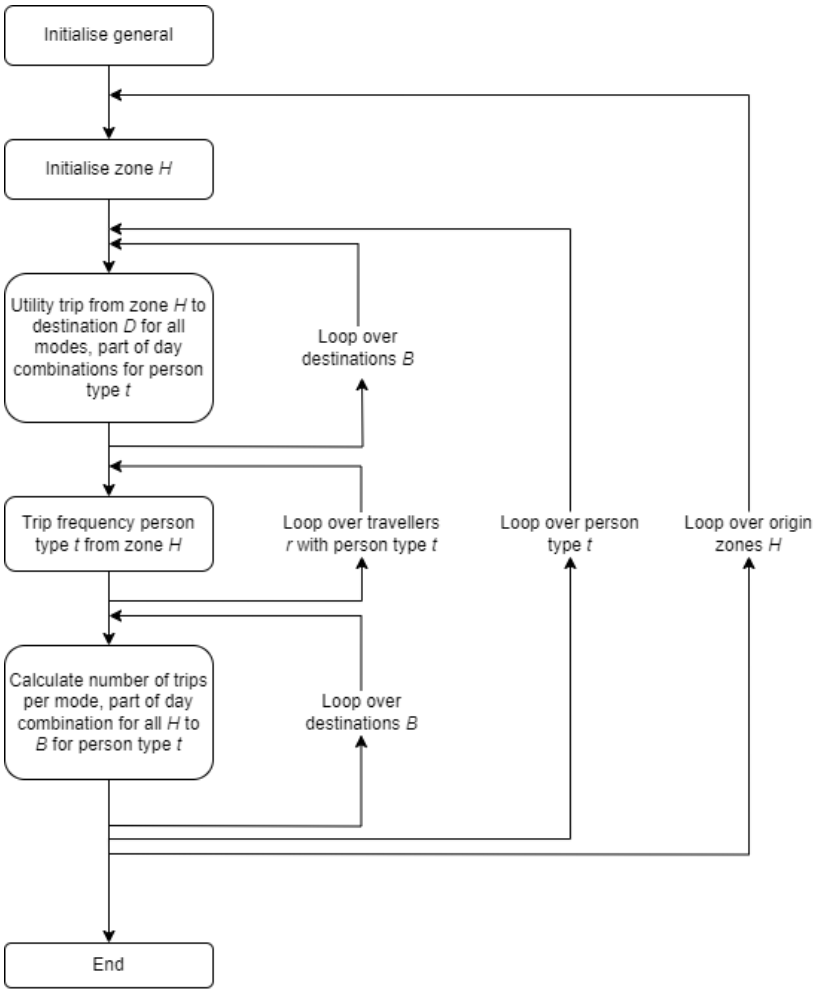

Figure 3.1: Detailed overview of the structure of SES. Figure translated from figure 3.1 from RWS WVL (2021g).

Input and initialisation

SES uses much different data as input, including the outputs of the previous modules, like car ownership coefficients. It also receives additional data, including zonal data (e.g. area, total population) or data about train stations (e.g. location, train frequencies). In some cases, some initial processing must be done before the input can be properly used in the model. For example, each zone needs to be coupled to one or more train stations to model train travel. The number of train stations that can be coupled to each zone depends on the DU, e.g. a zone with a DU of 1 gets coupled to a maximum

of 3 train stations and a zone with a DU of 5 or 6 gets coupled to maximum of 6 train stations. For each zone, the 'best' stations are chosen, which is based on the train frequency and the distance to the station.

For the initialisation, each household from the household sample[7], is coupled to one of the four categories of car ownership. After that all persons from the person sample[8] can get an additional characteristic based on the car ownership (e.g. no car and drivers licence; shared car). Finally, for each zone the chances of having a car are determined, which is based on household characteristics and zonal characteristics, as described in section 3.3.2.

3.3.5. Frequency (Module D7.1)

For each of the 12 travel motives the travel frequencies (e.g. home-work; home-shopping) are determined separately. (See table B.1 in the appendices for an overview of the motives.) To determine the frequencies a lot of data is used, including zonal data, accessibility data, the person and household samples and car ownership. The section is based on the documentation for module D7.1, RWS WVL (2021g) and module D10 (RWS WVL, 2021a). When other sources are used, it will be explicitly mentioned. An overview of the relevant variables and its corresponding values can be found in appendix B.

Each travel frequency model has the same two-step structure. First, the chance to take one or more tours is determined and in the second step it is determined if another tour will be made with the same motive, in the case a first tour was made. After each tour, the second step of the model will be applied. This is called a stop/repeat model. They use utility functions to determine the chance to take another trip.

Variables related to the spatial environment

The variables in the utility functions are based on eight different categories, including car ownership, DU and the logsum. All variables are dummy variables, and together they form the characteristics of the person. Not all variables are applied to each motive, nor to both steps of the model. The following variables are related to spatial characteristics:

- Five dummy variables for car ownership as described in section 3.3.4.
- Six dummy variables for the DU, where the first two dummy variables (DU is 1 or 2; DU is 3 or 4) apply to several motives, while the other 4 only apply to the motive home - business.
- The logsum, which was a measure for the accessibility of a zone and the type of person.

The travel frequencies are based on a sort of 'prototype' zone. The frequencies need to be increased for each individual origin zone to make them applicable for that zone. This is done based on the number of households in a zone of a certain type according to the household distribution (see section 3.3.1); the number of households in the sample; and the chance the household type has a car.

Effect of the spatial environment

When looking at the values, and especially the signs, of the different variables, it gives an estimation of the effect of the variable on travel frequency.

The different motives have different coefficient values for each variable, which means that the travel motive can affect how much the spatial environment affects the travel frequency. Living in an area with a low DU (1 or 2), generally has none or a negative effect on travel frequency for most motives. An exception is the chance of making a first tour for the motives home-work for fulltime workers and home-shopping. There, the low DU has a positive effect on travel frequency. A DU of 3 or 4 also has has a negative effect or no effect on travel frequency for most motives, except for the first tour a student makes with the motive home-education. When looking at the motive home-business, a DU of 1 and a DU of 2 has a negative effect on making a first tour. A DU of 1, 2, or 3 have a negative effect on making an additional trip. This same effect is seen with a DU of 4 or 5, though the negative effect on travel frequency is smaller.

[7] The household sample is based on stacked OViN data from 2015-2017 (RWS WVL, 2021a).
[8] The person sample is based on stacked OViN data from 2015-2017 (RWS WVL, 2021a).

There are no dummy variables for the the highest DU, which means that it serves as a reference. In other words, living in an area with a lower DU has for most motives a negative effect on travel frequency. The absolute value for most dummy variables increases (so the value itself decreases) when looking at a lower DU. So the expected outcome of the LMS is that the travel frequencies increase when a person lives in an area with a higher DU, when controlled for other variables.

To conclude, for determining the travel frequency, the included variables that are related to the spatial environment are limited. Most of the included variables have to do with personal characteristics (e.g. age group) or household characteristics (e.g. household size). There are several dummy variables for the DU (Density) and a variable for the accessibility (which is probably a combination of several D-variables, like Design, Destination accessibility and Distance to transit).

3.3.6. Mode, destination and part of day (Module D7.1)

In this part of the LMS, the utility of each possible mode, destination, part of day (MDD) alternative is determined. Based on these utilities, the different alternatives are divided over all available tours, as determined in the frequency model of the previous chapter. The probability of each choice combination is modelled using a nested logit model. See section 2.2 for more explanation about the structure of the model.

This section is based on the documentation for module D7.1 (RWS WVL, 2021g) and module D10 for the values of the variables (RWS WVL, 2021a). When other sources are used, it will be explicitly mentioned.

The MDD model consists of 6 nests: mode choice; mode choice for public transport; destination choice; part-of-day choice; access and egress mode choice; and station choice. Not every travel motive uses the exact same model structure or the same variables and not each nest is used for each mode (e.g. train uses all 6 nests, while walking only uses mode choice and destination choice). The variables used to describe each type of person can be divided into 10 categories. Several of these categories include spatial characteristics (e.g. zonal, size or station characteristics). The different categories are used in different levels of the nested logit structure, which also depends on the chosen mode. For example, the zonal characteristics are used in the destination choice when modelling cyclists and pedestrians, while it is included in the part-of-day choice for the other modes (car and public transport). This does not mean that the zonal characteristics for those modes are only relevant when choosing the part of the day. A variable affects the nest it is part of and all the 'higher' nests. For example, the zonal characteristics for train affect the mode choice, destination choice and the part of day choice. However, it does not affect the access and egress choices. See appendix C for a detailed overview of the nests.

In the following part of this section, an overview will be given of the different variables that are/ could be related to the spatial environment. Not all variables are applicable to each motive and to each mode. After that, the values (and especially the signs) of the corresponding coefficients will be analysed to see how that variable effects the MDD choice. To create a better overview, this will be done separately for each type of variable category. There are no relevant variables in the categories 'ASC', 'Household' and 'Part of day choice', so only 7 categories are given below. Interestingly, the coefficient for the motives 'child - shopping' and 'child - other' are the same for each variable.

See appendix B for an overview of the relevant variables and the corresponding coefficients.

Travel time

It is a bit dubious if a travel time variable can be counted as a variable related to the spatial environment. As shown in section 3.2, there is a relation between travel time and the spatial environment. The travel time variables for (e-)cycling and walking are part of the third level, the destination choice[9].

The travel time is in all cases negatively correlated with the choice of that alternative or in some cases there was no significant correlation. This is logical, because the longer it takes to travel to a destination, the less attractive this destination becomes.

[9] The documentation does not tell on which level travel time is used for the other modes. Presumably, this is done on the fourth level: 'part of day combinations'. This is based on the fact that travel time can depend on the part of the day for those modes (e.g. the frequency of public transport might be higher during peak hours) and because several other variables that are included on the third level for walking and cycling, are included on the fourth level for public transport and car.

Another variable in this category is the parking fee. This one applies to car drivers and all motives, except for the motive work-other and obviously the motives that only apply to children under 12. It is negatively correlated with the choice of an alternative.

Person

The Person category is part of the third level (destination choice) for (e-)cycling and walking and part of the fourth level (part of day choice) for the car and public transport choices. There are several dummy variables related to car ownership. See also section 3.3.2. They mostly apply to the modes car passenger and car driver, but (not) having a car also has an effect on train use and other forms of public transport for several motives. Car ownership is indirectly related to the spatial environment, as shown earlier in this chapter.

When looking at the coefficients, not having a car in the household has a positive effect on public transport choices for several motives. Generally, not having a car available or having to share a car, does have a negative effect on choosing the car.

There are no other variables in this category that are related to the spatial environment.

Size

The Size category is part of the third level (destination choice), which means that these variable play a part in both mode and destination choice. However, all variables in this category are applicable to each mode and the coefficients have the same value for each mode. In other words, they are really only relevant in the destination choice. The logarithm of these variables is taken before using them in the utility functions. The variables are explained below. Each variable in this category is related to the spatial environment.

- The total number of jobs in a zone, which is only relevant for the motive work-business. It is positively correlated with the choice to travel to that zone.
- The total number of jobs in the service industry which is relevant for the motives home-shopping and home-other. This variable is negatively correlated with the destination choice.
- The total number of jobs in the retail sector, which applies to the motives home-business, home other, child-shopping and child-other. This is positively related with the destination choice for all motives.
- The total number of jobs in the agricultural sector for the motive home-business, which also has a positive correlation with destination choice.
- The total number of places for students in special education, which applies to the motive child-education. This variable is negatively correlated with the destination choice, which also seems a bit of a contradiction.
- The population density in the destination zone for the motives home-other and all child related motives. This is negatively correlated with the destination choice. In other words, for the above mentioned motives, a higher population density at the destination makes the destination less attractive.
- The population density in the origin zone for the motive home-other. This is also negatively correlated with making the choice[10].

Zonal

The category Zonal is part of the third level (Destination choice) for (e-)cycling and walking and in the fourth level (part of day choice) for car and public transport. In this category there are many variables that are related to the spatial environment. Most of them are dummy variables for the DU.

[10] This variable seems a bit out of place, because the origin zone and the motive are already known and this variable is applicable to each mode. Perhaps this variable affects the logsum (see section 3.3.4), which affects the number of trips originating in a zone.

- There are seven distance coefficients[11] that are applicable for all modes, but only for a subset of people (e.g. distance coefficient for full-time worker; distance coefficient if the person is older 54). Those coefficients primarily apply to the motives home-work or home-education. For most of the variables, there is a negative correlation, meaning that a destination gets less attractive when it is farther.

- There are two dummy variables for a high job density (>75 jobs/ha). One of these applies to all modes for the motive home-work, home-business and home-other, while the second dummy variable applies only to car drivers with a home-work motive.

  The high job density is positively correlated with home-work and home-other motives and negatively correlated with home-business. The latter is an interesting result because the home-business motive is positively correlated with the number of jobs in the retail and agricultural sector in a zone. The second dummy variable for car drivers and a home-work motive is negatively correlated with the high job density, with a larger absolute value (-0.4314) than the dummy variable for the same motives for all modes (+0.1447). In other words, a high job density has a positive effect on the destination choice for all modes, except for car drivers. This is in line with the existing literature, where car travel reduces in areas with a higher job or population density.

- There are 3 dummy variables for the DU of the origin zone for car drivers for the motives home-work, home-other and home-shopping. There is also one for the train for the motive home-work and one for car passengers with motive home-other[12].

  All dummy variables for the car are negatively correlated with car use. (The lower DUs serve as a reference) When looking at the home-work motive, the absolute value for the dummy variable with a DU of 4 (-0.1281) is smaller than the one with a DU of 5 or 6 (-0.3533). In other words, the car has a lower chance of being chosen as a mode, the higher the DU of the origin is. Being a car passenger is also negative correlated with an origin DU of 5 or 6. The correlation between an origin zone with a DU of 5 or 6 and train use is positive. All these values are in line with the results found in the literature.

- There are 9 dummy variables for the DU of the destination zone, divided over car drivers, car passengers, train, walking and (e-)cycling. They apply to a lot of different motives. Similarly to the origin dummy variables, these variables again are for a DU of 4; 5 or 6; and 4, 5 or 6. So the the lower DUs (1-3) together serve as a reference.

  For car drivers, a higher DU means that the location has less chance of being chosen for almost all motives. There are no significant relationships found for the motives home-business and work-business. A possible explanation could be that a person has less freedom choosing their destination for business trips and all car costs are paid by their boss.

  For car passengers the DU has a similar effect as for the car drivers. A destination with a higher DU has a lower chance of being chosen for all motives except home-business, the work-bound motives and child-education. No significant relationship was found for these motives.

  For train travel a positive relationship was found for the motives home-business and home-other and a destination with a DU of 5 or 6. This was also found for walking only then for all motives except home-business, work-business and child-education. This is also in line with literature where a higher DU is associated with more train and walking. Interestingly, there is no dummy variable for the DU for BTM. The reason for this is unknown[13].

  Somewhat surprisingly, there is a negative correlation with travelling by bike to destinations with a DU of 5 or 6, for all home-bound motives except home-business. The literature review found that the share of both walking and biking is generally higher in more urban areas.

- A dummy variable for a travel distance larger than 80 km for all modes with a motive home-other. This correlation is positive, so a farther destination has a higher chance of being chosen.

[11] No definition is given for the distance coefficient, but it is assumed that it is positively correlated with the distance to a destination.

[12] The documentation says that this variable is for the destination zone, but based on the variable name and the location of the variable in the table, it is assumed that this variable is for the origin zone.

[13] It seems unlikely that there is no relationship between BTM and the DU. It is possible that relationship was already incorporated through another variable, making the variable for the DU insignificant. It is also possible that there is a mistake in the documentation.

- The share of higher educated work in a zone, for all modes and the motives home-work and home-business. This relation is positive, so a destination has a higher chance of being chosen if there is a higher share of higher educated work in a zone.
- Three variables for the share of higher educated work in a zone if the person is higher educated; medium educated; or lower educated for the motive home-work. In line with expectations, there is a positive correlation with being higher educated and choosing a destination with a high share of higher educated work. There is a small negative correlation if the person is medium educated and a larger negative correlation if the person is lower educated.

Train

The category Train is part of the fifth level (access and egress choice) of the nested logit model. Most variables in this category are related to personal characteristics (e.g. having a student OV-chipcard), the type of access/ egress mode (e.g. constant for using the bus as access mode) or a combination of those (e.g. Being a student and using a bike as access mode)[14]. It could be argued that having access to certain access and egress modes is related to the spatial environment. Modes like the tram or metro are mostly available in more urban areas. However, it is difficult to make any conclusions based on the available variables.

There is also a ratio of access and egress distance to the total distance between the origin and the destination zone. This variable applies to the modes BTM, bus/tram and bus. A higher ratio means that the access and egress distance to and from the train station is relatively long. This is negatively correlated with all relevant motives[15], except for the motives child-shopping and child-other. These are positively correlated with this longer access and egress distance. This means that for most motives a proportionally large distance of access and egress travel, makes an option less attractive.

Station choice

The category Station choice is part of the sixth level of the nested logit model, which is also called station choice. This category has again several variables related to the spatial environment. None of the variables are relevant to work-bound motives[16].

- A dummy variable for if the station has a DU of 6 for the motives home-business and home-other, when car passenger is the access mode[17]. This has a negative effect on the choice of the alternative. This implies that someone is less likely to travel as a car passenger to a station with a DU of 6. This is in line with the literature, where car use decreases in high density areas (e.g. Schwanen et al., 2002).
- Parking fee in origin zone. Applicable to all motives, except the work-bound ones for car passengers. For car drivers the child-motives are also excluded[18]. The parking fee is negatively correlated with the station choice. In other words, people are less likely to use the car as access mode when the parking fee is higher.
- Two variables related to the number of car parking/ bike parking places divided by the number of departing trains per hour for the origin station. A third variable gives the relation between bike parking places and the number of departing train on the destination station. They apply to most motives.

[14] Almost all variables in this Train category apply to other modes than train. Presumably these modes are the access and egress modes, but it is not explicitly stated.

[15] There is not enough data to model train travel for the motives work-other and child-education.

[16] The documentation stated that the motive child-education is not modelled for train. There are however variables for this motive, which are identical to the variables for the other child motives. There are no significant variables for the motive work-business, even though that motive is modelled. It is unclear if there is a mistake somewhere, or if all station choice variables are genuinely insignificant for that motive.

[17] The description of this variable says nothing about the the access mode and does not clarify if the DU relates to the station or the origin zone of the tour itself. Based on the variable name 'CpAccSUrb6', it is likely the variable applies to car passengers (Cp), because this is a common used abbreviation in the documentation. 'SUrb' could imply that the DU corresponds to the DU of the station. This is a logical assumption based on the category of the variable (station choice) and the abbreviations used in the documentation (e.g. 'OUrb' corresponds to the DU of the origin zone). The analysis will be done based on these assumptions.

[18] There are two variables with the same description. Based on the variable names, it is likely that apply to the different access modes. It is unclear if the origin zone relates to the zone of the station or the start of the tour.

All these variables are positively correlated with all motives, or there was so significant correlation. In other words, when comparing two train stations with a similar frequency, the train station with more bike and/or car parking places has a higher chance of being chosen.

- There are several variables related to the access and egress travel time to/ from the station with several modes. This is including a variable for the walking time to an access or egress BTM stop and it includes waiting time for the access/egress. These variables apply to all motives, except the work-bound ones. All these variables are negatively correlated with the access/egress time. The shorter the travel is to the station, the higher the chance a person chooses that station.
- There are several variables related to the in-vehicle time in the train and penalties for transfers. They apply to all motives, except the work-bound ones. Both the in-vehicle train time and the penalties are negatively correlated with the station choice. Having to transfer, negatively impacts the station choice and by extent the choice to travel by train.
- Two variables for the job density in the retail sector. They apply to all motives except the work-bound ones[19]. The job density for the retail sector in both the origin station and the destination station is for all motives positively correlated with the station choice. So a higher retail job density around a station increases the chance of people using that station.

Conclusion

When looking at the variables that are used in the MDD choice, many of them are related to the spatial environment. However there is not a lot of variation in the type of variables. Most of the variables are part of the D-variable Density: All the dummy variables about the DU and the variables for the population density and the job density. The D-variable Distance to transit has also many different variables. The Train category has a variable related to the access and egress distance, which quite literally relates to the Distance to transit. The Station choice category also has several variables related to the access and egress travel time, which is also related to the Distance to transit. Especially when including the variable for walking time to a BTM access/egress stop. There is a variable for the waiting time for the access/egress which says something about the frequency and quality of the public transport. This can also be counted as Distance to transit. Besides that, there are several constants for different access and egress modes. It should be noted that the Distance to transit variables are primarily related to train use.

Several of the variables in the Size category could be counted as Destination accessibility. These are the variables that give the total number of jobs in different sectors and the number of places in special education. A higher number, means that these locations are better accessible. However, these only apply to a few motives. For example, the motive home-shopping has only the variable for the number of service jobs, which gives only a limited representation of Destination accessibility. The Zonal category also has variables for the share of higher educated jobs in a zone, which could also be counted as Diversity. Finally, there are some variables related to Demand management, namely the parking fee variables.

All in all, the different D-variables are mostly represented in the different variables in the MDD choice model. However, this representation is not equally. Where there are a lot of variables related to Density, the variables for Diversity are a lot less. Even when looking at the DU, the degrees are mostly grouped together in a single variable and there are no variables separating the lower three degrees, because they all serve as a reference together. See table 3.1 for an overview of the different D-variables found in the LMS documentation.

3.3.7. Additional destinations (Module D7.2 & D7.3)

After modelling the tours, secondary and tertiary destinations are modelled. This is done in the modules D7.2 and D7.3. There are no notable characteristics of the spatial environment implemented in these modules. From zonal data, some data about the number of jobs and inhabitants are used and variables like travel time. (RWS WVL, 2021h; RWS WVL, 2021i)

[19] Judging from the name of the variable, one variable is for the access station and the other from the egress station. This is not stated explicitly in the description.

3.3.8. Conclusions

All in all, there are many variables in the LMS that include the spatial environment to some extent. However, many variables only include a limited number of motives or modes, making the total number of variables that apply to each individual MDD choice considerably less. Besides that, a lot of variables are related to Density, but other aspects of the spatial environment are only included to a limited extent. It is questionable if the variables that are currently included are able to capture the spatial environment well enough.

It is important to keep in mind that the variables related to the Demographics (which were not covered in this section) are also indirectly related to the spatial environment. As seen in the Population module D4.1, the distribution of the population is also affected by Density variables, giving different kind of populations in different zones. The same can be said for car ownership.

In the next section, a clear overview will be given for the different D-variables and how they relate to the variables found in the LMS.

3.4. Comparisons LMS and literature

Table 3.1 gives an overview of the different D-variables with several examples how these D-variables can be quantified based on the literature. It also shows the different variables found in the LMS, sorted per D-variable and the number of different variants of that variable in the LMS. It is important to note that the number of times a variable appeared can be a bit misleading, because sometimes a variable only counted for a specific mode or motive. So in reality, not each variable would appear in each possible choice. Besides that, most of the variables are related to the destination zone, though there are several variables that are related to the origin. The counted variables come from the car ownership model, frequency model and MDD model.

First of all, there are a lot of different variables for Density and Distance to transit[20]. For Density, the LMS uses both population density and job density variables belonging to a zone, but also the densities including surrounding zones in a different radius. The DU also includes the population density of surrounding zones. This way, both larger dense areas can be identified, but also smaller 'peak' density zones. When comparing the Density variables from the LMS with those from the literature, it seems that this D-variable is well implemented in the LMS.

When looking at Diversity, the number of variables in the LMS seems to be a lot lower. There are a few variables that give the share of certain types of jobs. These variables can probably be put under Diversity (similarly to jobs-to-workers ratio), though other aspects of Diversity like land use ratios or the historical development of the city are not included in the LMS.

For Design, no explicit variables could be found in the LMS. It could be argued, however, that Design is incorporated in the output of the accessibility variable. Variables like the road density affect the travel distances and can make it more attractive to travel by certain modes.

For the variables that could be counted as Destination accessibility, the LMS again mainly includes job-related variables and one variable related to education. There are no variables related to the location of the zone with respect to the city centre or the average distance to several points of interest (other than jobs and education).

There are many variables included in the LMS that belong to Distance to transit, both variables related to the distances to stops and the quality of public transport (e.g. frequencies, parking places). The Distance to transit D-variable seems to be well implemented in the LMS. However, it should be noted that most Distance to transit variables are only relevant for train use, including access and egress, and not for tours were BTM is the main mode.

For Demand management, there are also several variables (parking fares and permits), which are similar to variables found in the literature.

[20] Note: it is unclear what the 16 access and egress constants for the different modes mean. It is assumed that they are related to the quality or accessibility of that access or egress mode in that zone, because not each mode is available in each zone (e.g. tram/metro).

Table 3.1: Summary of D-variables with examples found in literature and the D-variables found in the LMS. The final column shows how many different variants of that variable can be found in the LMS documentation (RWS WVL, 2021g).

D-variables	Examples from literature	Variables in the LMS	LMS counts
Density	Population density Job density Address density Proximity index (Puylaert et al., 2022)	Degree of urbanisation Population density (including surrounding area) Job density (including surrounding area) Job density detail sector DU in combination with demographic variable	22 3 4 2 2
Diversity	Entropy measure (Kockelman, 1997) Land use diversity indices (Limtanakool et al., 2006; Harts et al., 1999) Historical development city (Van De Coevering & Schwanen, 2006) Job to workers ratio Ewing and Cervero, 2010	Ratio of agriculture jobs Share of high educated work	1 4
Design	Road network density (Sung & Eom, 2024) Road ratio (Sung & Eom, 2024) Road width (Sung & Eom, 2024) Intersection density (Ewing & Hamidi, 2015) Height-width ratio (L. Liu et al., 2023) Continuity street wall (L. Liu et al., 2023)		
Destination accessibility	Distance to city centre (Næss et al., 2017) Average distance to points of interest (Thao & Ohnmacht, 2020)	Number of jobs in several sectors Number of student places in special education	4 1
Distance to transit	Distance to transit stops (Ewing & Cervero, 2010) Frequency of transit (Kent et al., 2023) Quality network (Kent et al., 2023) Development transit network (Sung & Eom, 2024)	Distance (to and of) access and egress for several modes Ratio parking places station bike/car to frequency station Access and egress constants Transfer penalties and waiting times	9 3 16 5
Demand management	Parking restriction (Kent et al., 2023) Parking fare (Kent et al., 2023) Commuter allowance (Sung & Eom, 2024)	Parking fare Parking permits	4 2
other		logsum: measure of accessibility of a zone Distance coefficients Accessibility of OD pairs	1 7

Finally, there are some variables included in the LMS that are difficult to place within a D-variable. The logsum is a measure of the accessibility of a zone, based on the utilities from the MDD choice model, which determines the number of trips originating from a zone. This means that the logsum is a combination of all D-variables (including Demographics) and is not part of a category. As said before, the accessibility is presumably related to the Design variable. It is unknown what the distance coefficients exactly entail, but they are presumably also related to Design or Destination accessibility.

To conclude, the D-variable framework, as obtained form the literature, is partly implemented in the LMS. Some variables (Density and Distance to transit) are implemented using many different variables. It is noticeable that the variables belonging to Diversity and Destination accessibility are primarily related to jobs. Other aspects of the spatial environment, like land use diversity or accessibility to other points of interest like recreation, health or different kinds of education are not or barely included.

Based on these findings it can be expected that the LMS will perform relatively well in modelling train travel in different regions. The other modes will presumably perform worse, due to the limited number of variables. From the remaining variables, most of them are related to Density. It is expected that modes that are shown to depend a lot on the population density, are better modelled than modes that are more dependent on other (D-)variables. Which modes that are will be discovered in the data analysis, next chapter. In the discussion, section 5.2.4.

Another point of interest is the distribution of the variables over the modes. As said before, most of the variables related to the spatial environment are for train use. The modes car driver and then car passenger follow after that. For BTM, there are no D-variables exclusively related to BTM, and for walking and cycling there is only one dummy DU-variable. There are, however, variables for travel time and distance coefficients for those specific modes and D-variables that apply to all modes (e.g. the variables from Destination accessibility).

This chapter provided the information to answer the first two sub-questions. Aspects of the spatial environment that affect travel behaviour were identified using the D-variable framework (sub-question 1) and it was discovered how aspects of the spatial environment were captured in the LMS (sub-question 2).

4

Modal split analysis for different spatial environments

This chapter shows the process and the results of the data analysis and provides the information to answer the last three sub-questions.

The first section of this chapter covers the filtering and processing of the data. After that, an exploratory data analysis is done to look for differences between the different DUs and interesting trends that can be discovered. Next, a cluster analysis is done with the goal to identify different type of regions based on spatial characteristics, with large differences in travel behaviour between the regions. Finally, propensity score matching is used to quantify the effect of the spatial environment on the modal split.

4.1. Filtering and processing the data

This section covers the filtering and processing of the data. This has been partly covered in the methodology (section 2.5.1). However, not all choices could be described in the methodology, because the literature provided additional input for the data analysis (e.g. the D-variables).

4.1.1. Match PC4, neighbourhood and LMS zoning

The LMS zones needed to be coupled with the PC4 and neighbourhood zones. To match the PC4 zones, a data file was provided that matched each LMS zone with a PC4 zone (RWS WVL, 2017). In total, 37 PC4s did not appear in this data file and 50 PC4s were matched to two different LMS zones instead of one. For those double PC4s, they were matched with only one of the LMS zones. This created possible uncertainties, because some OViN trips would be matched to a wrong (neighbouring) zone. The 37 PC4s without any match, were matched using a simple algorithm: the coordinates of the centroid of each PC4 zones were matched with the coordinates of the area of the LMS zone (i.e. the PC4 zone was matched with the LMS zone the centroid belonged to). This gave a good match for all the missing PC4s, except for 4. These postal codes did appear in the OViN dataset, but not in the PC4 dataset from 2019 that provided the coordinates of each PC4 zone (CBS & ESRI Nederland, 2019). These PC4s were presumably abolished in earlier years or a mistake was made in the OViN data[1]. Trips with one of these 4 PC4s were removed from the OViN dataset.

After matching all the trips to their corresponding LMS zone and removing the trips that could not be matched, the final version of the stacked OViN file contained 379,797 trips which were made by 115,396 individual persons. This gives an average of 270 trips departing per LMS zone. The maximum number of trips departing from a zone is 1930 and there are a few zones with 0 trips. 97% of the zones have more than 20 departing trips and 86% of the zones more than 100 trips.

Data for land use existed only on neighbourhood level from 2017 and not on PC4 level (Centraal Bureau voor de Statistiek & Kadaster [CBS & Kadaster], 2019). The neighbourhood zones were matched

[1] Older PC4 datasets were searched. 2 of the PC4s existed around 2012. No record of the other PC4s was found.

Table 4.1: Overview of all the variables used to create the clusters and the corresponding source.

Variable name	Description	D-variable	Source
Pop_dens	Population density [people/ha]	Density	RWS WVL, 2020
Surrounding_pop_dens	Population density of all zones in a radius of 3 km [people/ha]	Density	RWS WVL, 2020
DU	Degree of urbanisation	Density	CBS & ESRI Nederland, 2019
Job_dens	Job density [jobs/ha]	Density	RWS WVL, 2020
Surrounding_job_dens	Job density of all zones in a radius of 3 km [jobs/ha]	Density	RWS WVL, 2020
Residential	Ratio of landuse used for residential	Diversity	CBS, 2022a; CBS & Kadaster, 2019
Services	Ratio of landuse used for services	Diversity	CBS, 2022a; CBS & Kadaster, 2019
Industrial	Ratio of landuse used for industry	Diversity	CBS, 2022a; CBS & Kadaster, 2019
Nature	Ratio of landuse used for nature	Diversity	CBS, 2022a; CBS & Kadaster, 2019
Agricultural	Ratio of landuse used for agriculture	Diversity	CBS, 2022a; CBS & Kadaster, 2019
Infra	Ratio of landuse used for infrastructure	Diversity	CBS, 2022a; CBS & Kadaster, 2019
Nature_Agri	Ratio of landuse used for agriculture or nature	Diversity	CBS & ESRI Nederland, 2019
Entropy	Entropy measure	Diversity	CBS, 2022a; CBS & Kadaster, 2019
Special	National specialisation index	Diversity	CBS, 2022a; CBS & Kadaster, 2019
House_45_less	Ratio of houses built before 1945	Diversity	CBS & ESRI Nederland, 2019
House_45_75	Ratio of houses built between 1945 and 1975	Diversity	CBS & ESRI Nederland, 2019
House_75_05	Ratio of houses built between 1975 and 2005	Diversity	CBS & ESRI Nederland, 2019
House_05_more	Ratio of houses built after 2005	Diversity	CBS & ESRI Nederland, 2019
Job-workers ratio	Working population / number of jobs	Diversity	RWS WVL, 2020
Road_density	Length of road per area [km/ km2]	Design	Rijkswaterstaat [RWS], 2022d
Road_width	Average road width [m]	Design	RWS, 2022e; RWS, 2022b
Bike_walk_percentage	Share of road meant for walking or cycling	Design	RWS, 2022e; RWS, 2022c
Dist_to_center	Average distance to city centre [km]	Destination accessibility	RWS WVL, 2020
Dist_food	Minimum distance to several food related locations [km]	Destination accessibility	CBS & ESRI Nederland, 2019
Dist_commercial	Minimum distance to several commercial related locations [km]	Destination accessibility	CBS & ESRI Nederland, 2019
Dist_health	Minimum distance to several health related locations [km]	Destination accessibility	CBS & ESRI Nederland, 2019
Dist_recreation	Minimum distance to several recreation locations [km]	Destination accessibility	CBS & ESRI Nederland, 2019
Dist_education	Minimum distance to several education related locations [km]	Destination accessibility	CBS & ESRI Nederland, 2019
Dist_point_of_interest	Minimum distance to several points of interest [km]	Destination accessibility	CBS & ESRI Nederland, 2019
Distance_station	Distance to closest train station [km]	Distance to transit	RWS WVL, 2018b
Distance_ic_station	Distance to closest intercity train station [km]	Distance to transit	RWS WVL, 2018b
Freq_station	Train frequency of closest train station [trains/hour]	Distance to transit	RWS WVL, 2018b
Freq_ic_station	Train frequency of closest intercity train station [trains/hour]	Distance to transit	RWS WVL, 2018b
Distance_BTM	Average distance to a BTM stop [km]	Distance to transit	University of Groningen Geodienst, 2021
Distance_bus	Average distance to a bus stop [km]	Distance to transit	University of Groningen Geodienst, 2021
Distance_metro	Average distance to a metro stop [km]	Distance to transit	University of Groningen Geodienst, 2021
Distance_tram	Average distance to a tram stop [km]	Distance to transit	University of Groningen Geodienst, 2021
btm_lines	Number of different lines of closest stops	Distance to transit	University of Groningen Geodienst, 2021
Bus_lines	Number of different bus lines of closest stops	Distance to transit	University of Groningen Geodienst, 2021
Metro_lines	Number of different metro lines of closest stops	Distance to transit	University of Groningen Geodienst, 2021
Tram_lines	Number of different tram lines of closest stops	Distance to transit	University of Groningen Geodienst, 2021
btm_stops	Number of BTM stops within a certain radius	Distance to transit	University of Groningen Geodienst, 2021
Bus_stops	Number of bus stops within a certain radius	Distance to transit	University of Groningen Geodienst, 2021
Metro_stops	Number of metro stops within a certain radius	Distance to transit	University of Groningen Geodienst, 2021
Tram_stops	Number of tram stops within a certain radius	Distance to transit	University of Groningen Geodienst, 2021
Distance_TM	Average distance to a tram or metro stop [km]	Distance to transit	University of Groningen Geodienst, 2021
TM_stops	Number of tram and metro stops within a certain radius	Distance to transit	University of Groningen Geodienst, 2021
Parking_fare	Average parking fee [eurocents]	Demand management	RWS WVL, 2020
Road_parking	Parking area next to roads / total population [m^2/person]	Demand management	RWS, 2022e; RWS, 2022a

with the LMS in a similar way the missing PC4s were matched. This resulted in several neighbourhoods that were matched to 2 LMS zones. In a few cases, it was obvious to which zone the neighbourhood belonged. For the rest, it was assumed that each neighbourhood belonged fully to both LMS zone in the calculations. After matching, there were 2 LMS zones without any land-use data. For these LMS zones, the nearest neighbourhood zones were manually picked. For more details and figures about this matching process, see appendix D.

4.1.2. Data for D-variables

According to the literature, the D-variable framework can help quantifying the spatial environment. This subsection will present the different D-variables that were found in the literature and give the method that was used to quantify these variables. Each variable will be determined for the whole zone. Later, these D-variables can be used in the cluster analysis to create different clusters.

Table 4.1 gives an overview of all different variables that were collected for this thesis. This table includes the source of each variable. The variable names are often used in the different figures later in this chapter.

Density and degree of urbanisation

For the Density variable, the population density, the job density and the DU are used. For the job and population density, the density of the zone itself is taken and the density of each zone, including the area and population of zones where the centre is within a radius of 3 km of the zone. This is done in a

similar way as calculating the DU. See appendix E for additional figures of the Density variables

The calculation of the DU described in the LMS documentation (see section 2.2.2). As told in section 3.1, the CBS uses a different DU. A comparison between the LMS definition and the CBS definition is given in figure 4.1. (The DU according to the CBS was given in the PC4 dataset (CBS & ESRI Nederland, 2019). No additional calculation were needed.)

An interesting difference between the two DUs is that the LMS takes the population density of the surrounding zones into account (to a larger extent), which gives larger areas with the same DU. The highest degree is only seen in the three largest cities: Amsterdam, Rotterdam and The Hague. The majority of the Randstad seems to have a degree of 3 or higher , while outside the Randstad most zones seem to have a degree of 3 or lower, with the exception of a few cities like Eindhoven or Groningen. The CBS however, takes a way smaller region into account when calculating the DU. This gives a lot more small areas with a high DU, while directly neighbouring zones might have a significantly lower degree. Many cities seem to have the highest degree, even outside the Randstad.

Because the LMS uses the surrounding zones to calculate the population density and the DU, only cities that have both a high density and are large have a high DU. This is in contrast with the CBS where also smaller cities like Groningen or Eindhoven can have the 'highest' DU.

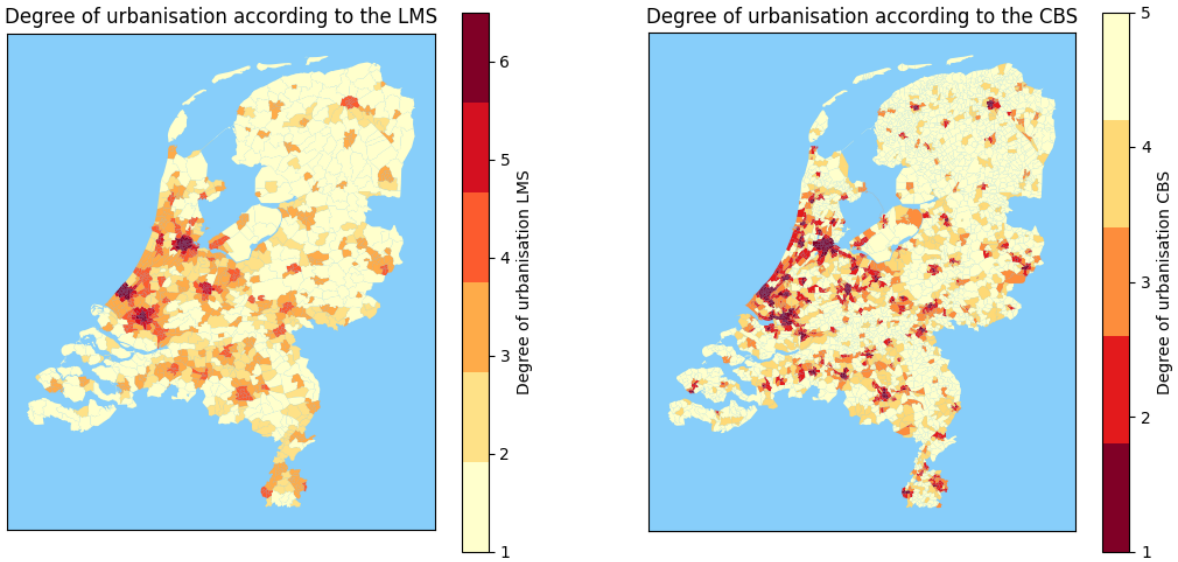

Figure 4.1: Comparison between the zones according to the LMS on LMS zone level (left) and the CBS on PC4 zone level (right) in the Netherlands. The maps are based on RWS WVL (2020) (left) and CBS & ESRI Nederland (2019) (right).

Diversity

The variable Diversity stands for the different land-uses in an area. This can be measured using indices (e.g. Limtanakool et al., 2006; Harts et al., 1999; Kockelman, 1997), the job-workers ratio (e.g. Ewing and Cervero, 2010; Næss et al., 2017) or the historical development of a city (Van De Coevering & Schwanen, 2006).

To determine the Diversity variables related to land-use, it is important to separate the land-use in different categories. Several sources were used as an inspiration to determine the land-use types (Limtanakool et al., 2006; Harts et al., 1999; Kockelman, 1997; Feng et al., 2013). The data that was used to calculate the land-use for each LMS zone is based on CBS (2022a) and CBS & Kadaster (2019)). These sources used 5-7 different land-use types to calculate the land-use balance and the other land-use indices. Due to the available data it was not feasible to directly copy one of the sources (e.g. some sources made a distinction between industrial and offices, while those were part of the same category in the available dataset). On top of that, a new land-use type was added for this thesis that was not used in any of the other source: agricultural. Agricultural land covers more than 50% of the Netherlands (CBS, 2022a; CBS & Kadaster, 2019) and does not fit well in the 'green & recreation'

category nor the ‘industrial & offices’ category, but might be more a combination of these two. Table 4.2 shows the land-use types that will be used in this thesis. See appendix E for all the Diversity figures.

Table 4.2: The six land-use categories that will be used to characterize the different regions. The total area and percentage of total area are based on (CBS, 2022a; CBS & Kadaster, 2019). The descriptions of the categories are based on CBS (2022b).

Category	**Area [ha]**	**Percentage**	**Description**
Residential	241,408	5.81 %	Homes, schools, small parks, gardens, playgrounds etc.
Services	40,548	0.98 %	Shops, restaurants, governmental buildings, police, hospitals, churches, etc.
Industrial & offices	122,865	2.96 %	Offices, factories, harbors, storage area, dumps, construction site, etc.
Recreation & nature	1,403,928	33.79 %	Parks, sports fields, camping sites, forests, lakes, etc.
Agricultural	2,230,445	53.69 %	Livestock, greenhouses, agriculture, scattered farms, etc.
Infrastructure	115,108	2.77 %	Railways, distributor roads, highways, gas stations, airport, etc.
Total	4,154,302	100 %	

In total, there are three land-use indices that are used for this thesis. The indices will be determined for each LMS zone. The proportion of each land-use type for each LMS zone is known. See 2.5 for more details about determining the land-use for each zone.

The first index is the local specialisation index (LSI), which is the proportion of each land-use type in each zone, see equation 4.1 (Limtanakool et al., 2006).

The second index is the national specialisation index (NSI). The version used in this thesis is based on Harts et al. (1999). This index measures how similar the land-use distribution of a zone is to the land-use distribution of the whole country, i.e. a zone of which the proportion of each land-use type is similar to the proportion of each land-use type in the whole country, has a low value and zones that are very ‘specialised’ have a higher value. To discover why a zone has a high NSI, it is useful to look at the LSIs of that zone. The NSI does not show which land-use type(s) are different. See equation 4.2.

Finally, the entropy measure or land-use balance. This index measures if all land-uses are equally represented in the zone and is used a lot in the literature (e.g. Kockelman, 1997; Limtanakool et al., 2006; Feng et al., 2013). Equation 4.3 shows how to calculate this balance. The index will vary between 0 and 1, where 1 means that there is a perfect balance between all types of land-use. It is important to note that the above mentioned sources calculate the entropy measure for ‘developed’ land use of a large urban area. In this thesis, the entropy measure is calculated for a whole country and all land is included. This could give different results than found in literature.

$$LSI_{jk} = \frac{A_{jk}}{A_k} = P_{jk} \tag{4.1}$$

$$NSI_k = \frac{1}{2} \sum_j \left| \frac{A_j}{A_{tot}} - P_{jk} \right| \tag{4.2}$$

$$Entropy_k = -\sum_j \frac{[P_{jk} \times ln(P_{jk})]}{ln(J)} \tag{4.3}$$

Here, j is the land use type ($j = 1, 2, ..., J$); k the LMS zone ($k = 1, 2, ..., K$); P_{jk} the proportion of land-use type j in zone k (i.e. the LSI) and A the area of a certain zone and/or land-use type (A_{tot} is the area of the whole country).

The job-workers ratio is very straightforward, see equation 4.4. Here, $n_{workpop}$ is the total working population in a zone (people that work or want to work 12 or more hours per week) and n_{jobs} the total number of jobs in a zone. This variable is based on RWS WVL (2020).

$$ratio = \frac{n_{workpop}}{n_{jobs}} \tag{4.4}$$

The historical development of a zone is based on the built year of its houses. Limtanakool et al. (2006) uses 3 categories: Houses built before 1945, houses built between 1945 and 1970 and houses built after 1970. Due to the available data and the fact that this source is from almost 20 years ago, it was decided to add an extra category for newer houses. This gives the following list:

- Houses built before 1945
- Houses built between 1945-1975
- Houses built between 1975-2005
- Houses built after 2005

For each LMS zone, the percentage of houses of each category was determined on PC4 level and aggregated to the LMS zones, in a similar way as equation 2.4. These variables are based on CBS & ESRI Nederland (2019). The original CBS dataset contained the number of houses built in categories of 10 years. However, due to privacy reasons all categories with less than 5 houses are censored (Van Leeuwen & Venema, 2023). Because of this, the final dataset contains several missing values.

Design

The Design variable says something about the characteristics of the street network. There are many ways to do this. Based on the available data and what was deemed relevant for the Netherlands, the road network density (Sung and Eom, 2024; Li et al., 2024), the road width (Sung & Eom, 2024) and the proportion of bike/pedestrian paths (Ewing & Cervero, 2010) are used. Variables like the intersection density (Ewing & Hamidi, 2015) are presumably more relevant for more grid-like and car centered road networks like the US and variables like the height-width ratio (L. Liu et al., 2023) of the road require more data than is available for this thesis.

To calculate these variables, shapefiles from the national road database (Nationaal Wegenbestand [NWB]) are used (RWS, 2022d; RWS, 2022e; RWS, 2022b; RWS, 2022c; RWS, 2022a). To use this data, each road had to be matched to its corresponding LMS zone. This was done using a *GeoPandas* function, similarly to the method used to match the PC4 and neighbourhood zones with the LMS zones. A few sample zones were taken to check if the roads were matched to the right LMS zones, see appendix D. The NWB shapefile could be matched with additional NWB files that contained more information about the roads, like the type of road or the road width.

The road network density of a zone was calculated using equation 4.5. Here l is the length [km] of all roads in a zone, so not only highways or large roads, but also the smaller roads within neighbourhoods or specific bike/pedestrian paths. A is the total area of the zone [km^2].

$$density_{road} = \frac{l}{A} \tag{4.5}$$

The average road width is calculated by taking the average of the width of each road section in a zone. However, due to missing data the width of not all roads is given, which might make this statistic less reliable, especially if there is a bias that smaller of wider roads were not measured.

The proportion of bike and pedestrian paths is calculated by dividing the total length of these roads by the total length of all roads in a zone. It is important to note however, that the dataset for calculating this proportion is from 2022 and not 2018. In the past years, many new bike paths were created and the dataset has still not included all bike paths. Besides that, it only takes separated bike paths into account, so a car road with an unprotected bicycle lane next to it is not counted as a bike path. This might make this variable less reliable (Nationaal Wegenbestand [NWB], 2021). See appendix E for maps of the different Design variables in the Netherlands.

Destination accessibility

The Destination accessibility variable says something about how easy different trip destinations can be accessed. For this the distance to the city centre (e.g. Næss et al., 2017; Næss, 2006; Ewing and Hamidi, 2015; T. Liu and Ding, 2024) was used and the average distance to various points of interest (Thao & Ohnmacht, 2020).

According to the literature, the distance to the city centre seemed to be an important variable when looking at travel behaviour. However, none of the sources gave an exact definition of what is counted as a city centre. Of the reviewed literature, only Næss et al. (2017) noted that the city centre (or downtown) is a location with many jobs concentrated. This could be due the fact that many studies are focused on a smaller area with one or a few clear centres, that could be selected manually. However, this is not the case in this thesis, so a simple method was developed to determine the location of the city centres.

Looking at the definition from Næss et al. (2017), it seems that not each shopping centre or central location of a city or village is counted as a city centre, but city centres are larger locations that attract a lot of people. However, it would be unfair to only look at absolute attraction values (e.g. job/ population density), because then the biggest part of the Randstad would be marked as a city centre and the rest of the Netherlands would have no centres. To mitigate this problem, both absolute and relative measures were used to determine the city centre.

First, a measure is needed to quantify the attractiveness each zone. For this, the population density and the job density were added to create a new density variable, see figure 4.2.a. After that, all zones with a combined density higher than 20 units/ha were selected[2]. This value was chosen because it seems to give a good balance between enough potential city centres outside the the Randstad, while not giving each rural area its own city centre, see figure 4.2.b. As shown in the figure, most of the Randstad was still marked as a city centre.

For the next step, it was decided that each municipality could have a maximum of 1 city centre; the zone with the highest combined density. By limiting the number of city centres in an area, the ‘relative’ city centre of an area can be found. This gave a total of 131 centre zones in the Netherlands, see figure 4.2.c. Finally, the distance to each city centre was determined by calculating the euclidean distance between the centroid of a zone to the centroid of the nearest ‘city centre’ zone, see figure 4.2.d. As shown in the figure, the final measure for the distance to the city centre still shows that people in the Randstad live on average closer to a city centre, than people outside the Randstad, which is in line with expectations. However, in the rest of the Netherlands there are still a large number of city centres left. A possible shortcoming of this method is that it only takes city centres in the Netherlands into account. People living close to the border, often have a large distance to the nearest city centre, while there might be a closer city centre in a neighbouring country.

The next variable is the average distance to a point of interest. Yu and Higgins (2024) researched 15-minute cities[3] and determined based on a literature review what kind of necessities should be present in such a city. At the end, they distinguished 5 different categories of necessities that people need. These categories will be used in this thesis as a way to determine the different points of interest. The PC4 dataset (CBS & ESRI Nederland, 2019) contained the minimum distance using the roads from each PC4 area to a lot of different locations. These locations were sorted into the 5 categories below and the average distance was calculated for all locations in the category and aggregated to LMS zone level, using an equation similar to equation 2.4.

- **Food**: grocery stores, restaurants, etc.
- **Commercial**: Department store, specialty store, etc.
- **Health**: Doctors, pharmacy. hospital, etc.
- **Recreation**: Library, swimming pool, theater, cinema, etc.
- **Education**: Child care, primary school, high school, etc.

[2] units is defined as the sum of the number of jobs and the number of inhabitants

[3] 15-minute cities is an urban planning concept, that stimulates denser cities, high land use and active modes. See Yu and Higgins (2024) for more information.

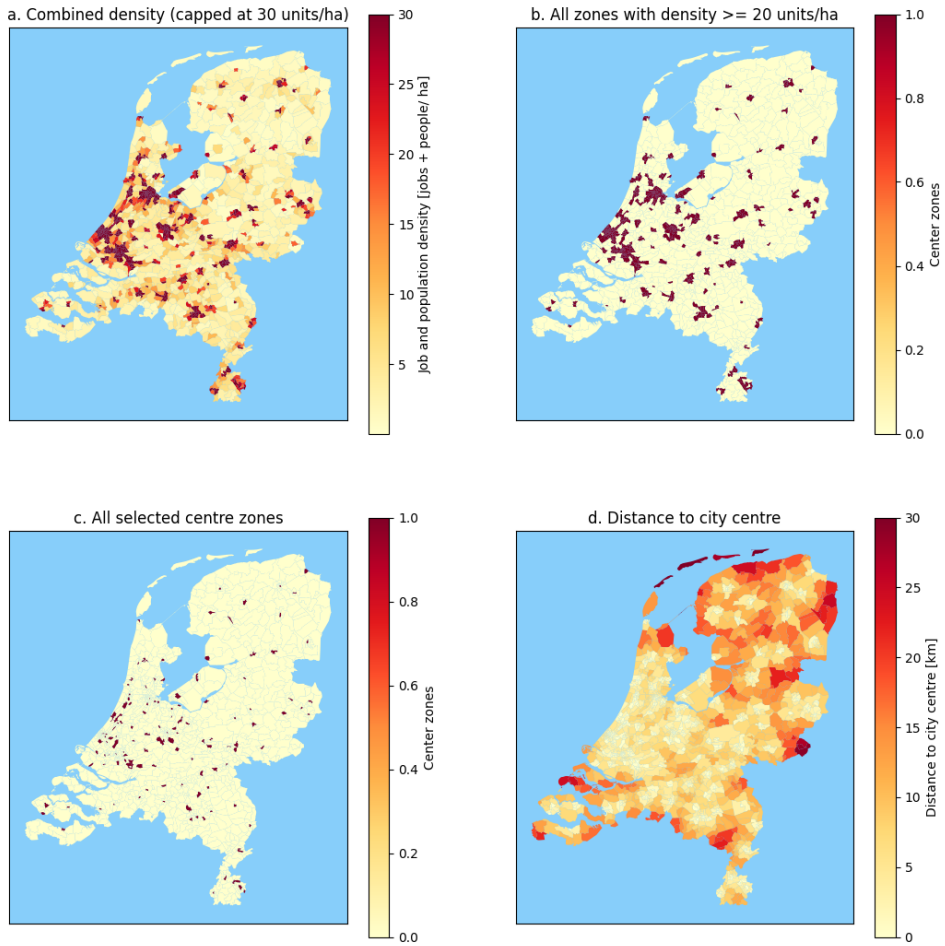

Figure 4.2: Methodology of finding the distance to the city centre in the Netherlands. The map is based on RWS WVL (2020).

Besides the average distance to those 5 categories, the average distance of all the categories was taken to get one general distance to a point of interest for each zone, see figure 4.3. This map shops that Destination accessibility in the Netherlands is on average fairly high, though it is, generally, the best in the (large) cities and the Randstad. Zeeland, Friesland, Groningen en Zeeland have outlier zones that have a large distance to points of interest. Again, it is important to note that points of interest and roads outside the Netherlands are not taken into account. This means that zones close to the border might show larger distances than the inhabitants travel in reality (CBS, 2012). See appendix E for maps of the 5 different categories and additional information.

Distance to transit

The Distance to transit variable gives an indication of the quality and accessibility of public transport of that zone. For this variable the the minimum distance to different kind of public transport stops was taken, the frequency of the train stations, number of different bus, tram and metro (BTM) lines and the number of BTM stops within a certain radius. This way not only the availability, but also the quality of the public transport network is taken into account (Kent et al., 2023). For the train stations, station data from the LMS (RWS WVL, 2018b) was used, while the BTM stops were assessed using data from University of Groningen Geodienst (2021).

First, the distance to the closest train station, the closest intercity train station and the closest BTM stops were taken. For this the euclidean distance was used between the centroid of a zone and the closest zone/ stop. For the BTM stops, the average distance to the 6 closest BTM stops was used. This was done because, contrary to train stations, most zones contain one or more BTM stops. So if there is a large rural zone that happens to have one bus stop at the center of the zone, it would score much

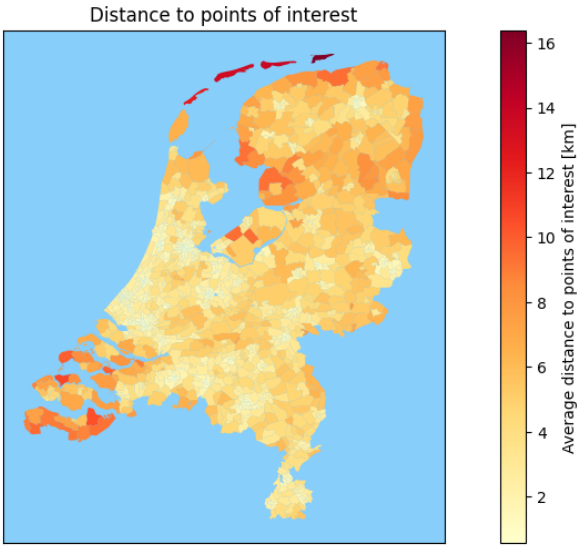

Figure 4.3: Average distance to several points of interest in the Netherlands. The map is based on CBS & ESRI Nederland (2019) and RWS WVL (2020)

higher compared to a similar zone with 3 bus stops that are located farther from the center. An even number was used because BTM stops in the dataset were counted in pairs (both directions separately), so by using an even value not only 'half' of a stop is taken into account. Only BTM stops within a radius of 5 km of a zone centroid are used. The distances were calculated using all BTM stops and for the bus, tram and metro individually. There are large parts of the country that do not have any access to the tram or metro. Zones that did not have any stops within a radius of 5 km got no stops assigned to them. When using a clustering method, it is not possible to have any missing values. To solve this, all zones without a BTM stop within a radius of 5 km, were given a value of 5 km.

A second measure for the accessibility of the BTM network was developed to measure the density of the network more accurately. The number of BTM stops within a 2.5 km radius of the centroid of a zone was counted, this was also done for bus, tram and metro individually. For this, stops that shared a same name and mode were removed from the dataset, making it an upgrade from the previous measure. It is assumed that zones with more BTM stops, have a BTM network that is more accessible.

For the train stations, the frequency was determined in trains per hour. This frequency sometimes differed between morning peak hours, evening peak hours and the rest of the day. The average frequency of those 3 values was taken as measure for frequency. This way the peak hour frequencies counted relatively more to this value than the frequency of the rest of the day. The BTM dataset did not contain any data about the frequencies. It did, however, contain data about the number of different directions served by that stop. This is a measure for how connected the stop is with the network. The 6 closest stops were used to calculate the average number of directions served by a BTM stop for each zone. See appendix E for maps of the Distance to transit variables.

Demand management

Travel Demand management are measures to stimulate or dissuade the use of certain modes. This thesis used the parking fee in a zone and a measure for parking places (Kent et al., 2023). Other measures like commuter allowance from employers (Sung & Eom, 2024) are not regulated nationally/ regionally, but differ between different companies. This means that it is not a suitable variable to use in this thesis.

The parking fee per zone is obtained from the LMS data (RWS WVL, 2020). For the parking places, the total parking area adjacent to roads can be obtained from the NWB from a dataset from 2018, in a similar way the Design variables were obtained. The total parking area was divided by the population of a zone to scale it. It is questionable how reliable this variable is. The dataset only counts parking places next to roads, so parking garages or other large parking lots are not taken into account. See

appendix E for maps of the Demand management variables.

Demography

Finally, the Demography. Even though the Demography is not a spatial environment factor, it is still important to control for it, because in different areas might live different kind of people. To decide which variables to include, eight studies from the literature review were evaluated to see which variables they took into account (Feng et al., 2013; Schwanen et al., 2002; Limtanakool et al., 2006; Rubin et al., 2014; Dargay and Hanly, 2003; Van De Coevering and Schwanen, 2006; Susilo and Maat, 2007; Van Acker et al., 2011).

Based on the characteristics from these studies, a list was made of characteristics to include. None of the studies included student OV (a free public transport card for students in the Netherlands), but it seemed like a good variable to include. The final list is given below. All demographic data is based on the combined OViN dataset. This is because the demographic data will not be used to compare zones, but to perform the propensity score matching as described in section 2.5.3, which uses the characteristics of individual people. The full list of characteristics is given below. These characteristics were determined for each person in OViN.

- **Age**
- **Gender**
- **Household income**: The standardised disposable household income class was used, instead of the income itself. This is the household income after taxes and certain mandatory insurances, standardized for a household size of 1 person. It is a measure for welfare (e.g. a person belongs to the 10% with the highest income).
- **Household size**
- **Household type**: 4 different household types were distinguished (1 adult; 2 or more adults; 2 parents with one or more kids (and possibly other people); 1 adults with one or more kids (and possibly other people)).
- **Social participation**: 4 different categories of social participation were distinguished (Part time worker; full time worker; student; other).
- **Education**: 6 types of education levels were distinguished, ranging from primary education to university. There is a separate category for people under 15.
- **The number of cars in the household**
- **Driver's licence**
- **Student OV**: The person has a student card for free public transport (OV) during the weekend or the week.

Approximately the same (average) demographic characteristics were also determined for the whole zone based on data from CBS & ESRI Nederland (2019) and RWS WVL (2020). With the exception for education, of which there was no data available. This data was used when doing tests for propensity score matching. See appendix E for Demography maps.

4.2. Exploratory data analysis results

An exploratory data analysis is to get an idea of the differences in travel behaviour between the different DUs, according to the LMS definition, and the differences between OViN and the LMS. This analysis focuses on the modal split. See section 5.1 for the reasons of this limited scope.

First, the modal split for the whole country separated by the DU is analysed. After that, several interesting areas will be analysed in more detail. At the end, an overview will be given of the most important results and conclusions of this section.

4.2.1. Modal split based on the degree of urbanisation

Figure 4.4 shows the comparison in modal split between OViN and LMS for the different DUs. The top row shows the percentage of trips that depart from a zone with a certain DU for all modes. The middle row shows the absolute difference between OViN and LMS, where a positive number means that the LMS overestimates the mode use. The bottom row shows the relative difference between OViN and LMS.

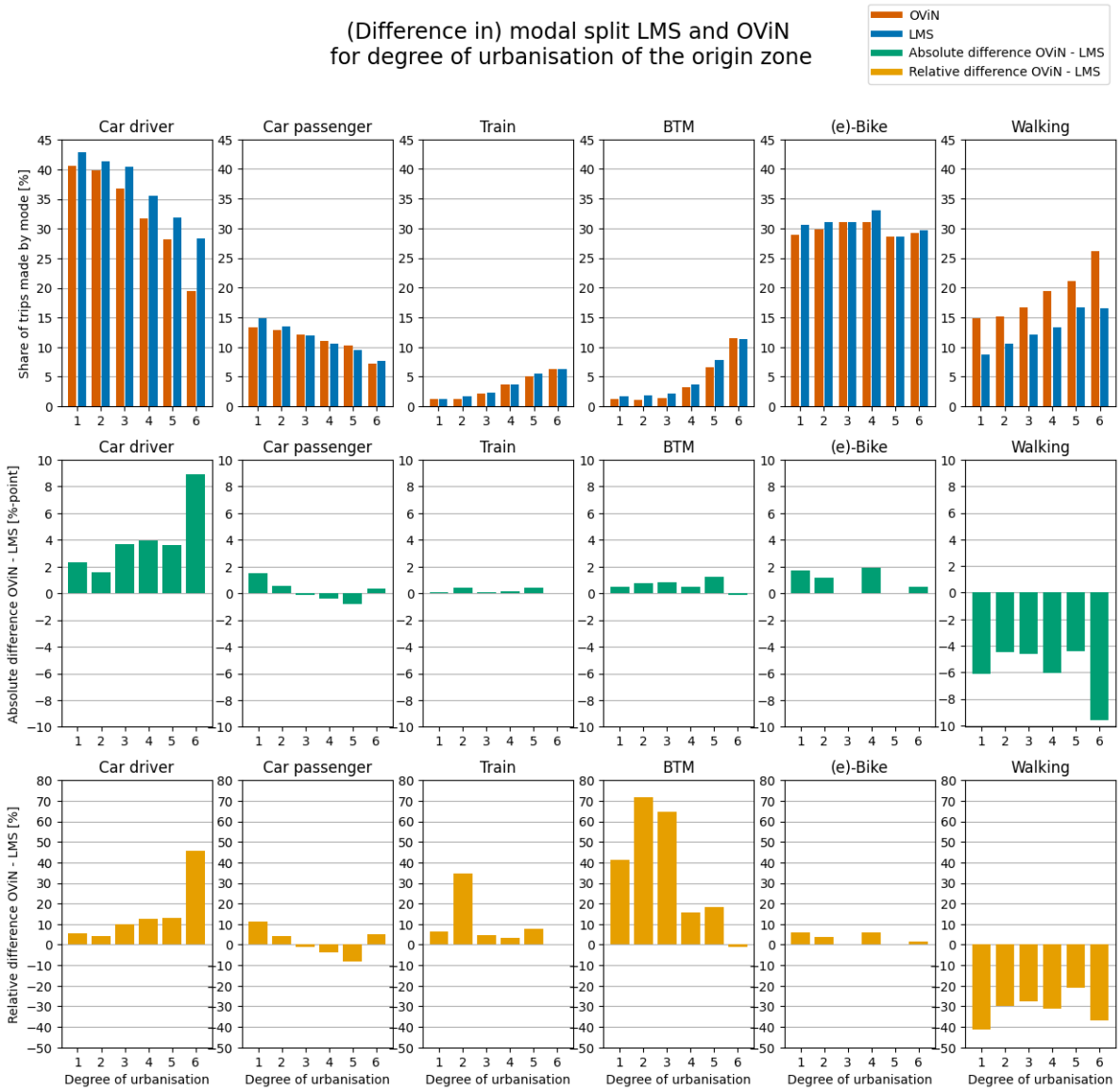

Figure 4.4: Top: Comparison between the modal split for different degrees and urbanisation and for OViN (red) and LMS (green). The percentages are based on the number of trips. Middle: Difference in percent points between OViN and LMS (OViN minus LMS). Bottom: Relative difference between OViN and LMS (difference = (LMS - OViN) / OViN). This graph is based on the combined OViN dataset from 2013-2017 and the LMS OD-matrices (RWS WVL, 2018c).

Both OViN and LMS generally seem to follow the same trend when the DU increases: car traffic decreases, public transport and walking increases and cycling stays approximately the same. This is mostly in line with expectations and findings from literature. It is also logical to see similar trends on national level because the LMS has been calibrated using OViN data.

However, the LMS seems to overestimate car driver trips, while underestimating walking trips. Especially when looking at the highest DU. The LMS predicts that around 28% of the trips is done by car drivers when the origin zone has a DU of 6, while OViN shows that less than 20% of the people take the

car in that case. The reverse is seen when looking at the number of trips done by foot. The predictions for the other modes seem to be a bit more accurate.

When looking at the relative differences, it shows that the relative difference is very high for BTM predictions with a low DU. However, due to a relatively small number of BTM trips, the relative error makes the differences seem very large. It is possible that the LMS is worse at modelling BTM trips with a low DU. However, the size of the OViN data set is also limited, making the chance higher that OViN does not give an accurate representation of smaller volumes of mode use.

Train use has also a relatively low number of trips. However, the relative differences (except for a DU of 2) are still lower than 10%. This implies that the LMS is better at modelling train use than BTM use.

Interestingly, bike use does not increase with the DU. It increases slightly with the first four DUs, but then decreases with a few percent points for the highest two DUs. This trend is correctly captured in the LMS and was already expected based on the LMS documentation. In section 3.3 it was discovered that there was a dummy variable for cycling to a destination with a DU of 5 or 6 with a negative value, indicating that people are less likely to take the bike to those zones, compared to the rest of the country. This can have a few causes. First, public transport is presumably better in higher density areas which can form a competition to bicycle use. Another explanation could be that shops and other necessities are closer in very dense areas, so people can do those trips by foot instead of taking the bike. Figure 4.3 shows that the average distance to points of interest in the Randstad (i.e. zones with a high DU) is generally lower than outside the Randstad.

All in all, the LMS captures trends in mode use relatively well when looking at the different DUs (i.e. the DUs with high and low use of each mode are captured well). However, there are still large differences in the 'absolute' share of trips, especially for car and walking.

After looking at the differences between the different DUs, it is also interesting to zoom in a bit on different areas to look differences on zone level. A few interesting locations that were found when looking through the data, are presented in the next sections. All zones with less than 20 trips recorded in OViN are removed from the maps, because it is assumed they contain too few data points to give a realistic view of travel behaviour in the zone.

4.2.2. Modal split Amsterdam

Figure 4.5 shows the DU in Amsterdam and figure 4.6 shows the modal split for Amsterdam according to OViN, LMS and the difference between the two. In the right column, a red zone means that LMS overestimats OViN and predicts higher values. Blue zones mean that the LMS predicts lower than seen in OViN. The difference shown here gives the absolute difference in mode share in %-points. For the relative differences, see appendix F.

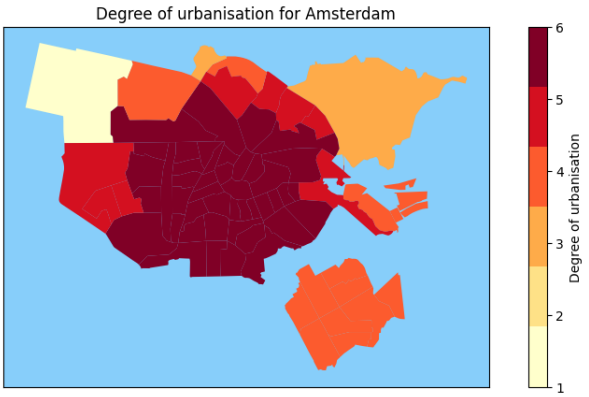

Figure 4.5: Degree of urbanisation for Amsterdam, according to the LMS. The map is based on (RWS WVL, 2020).

A large difference between the modal split according to OViN and according to LMS is that the LMS has less differences in travel behaviour between different zones, while the OViN shows a lot more

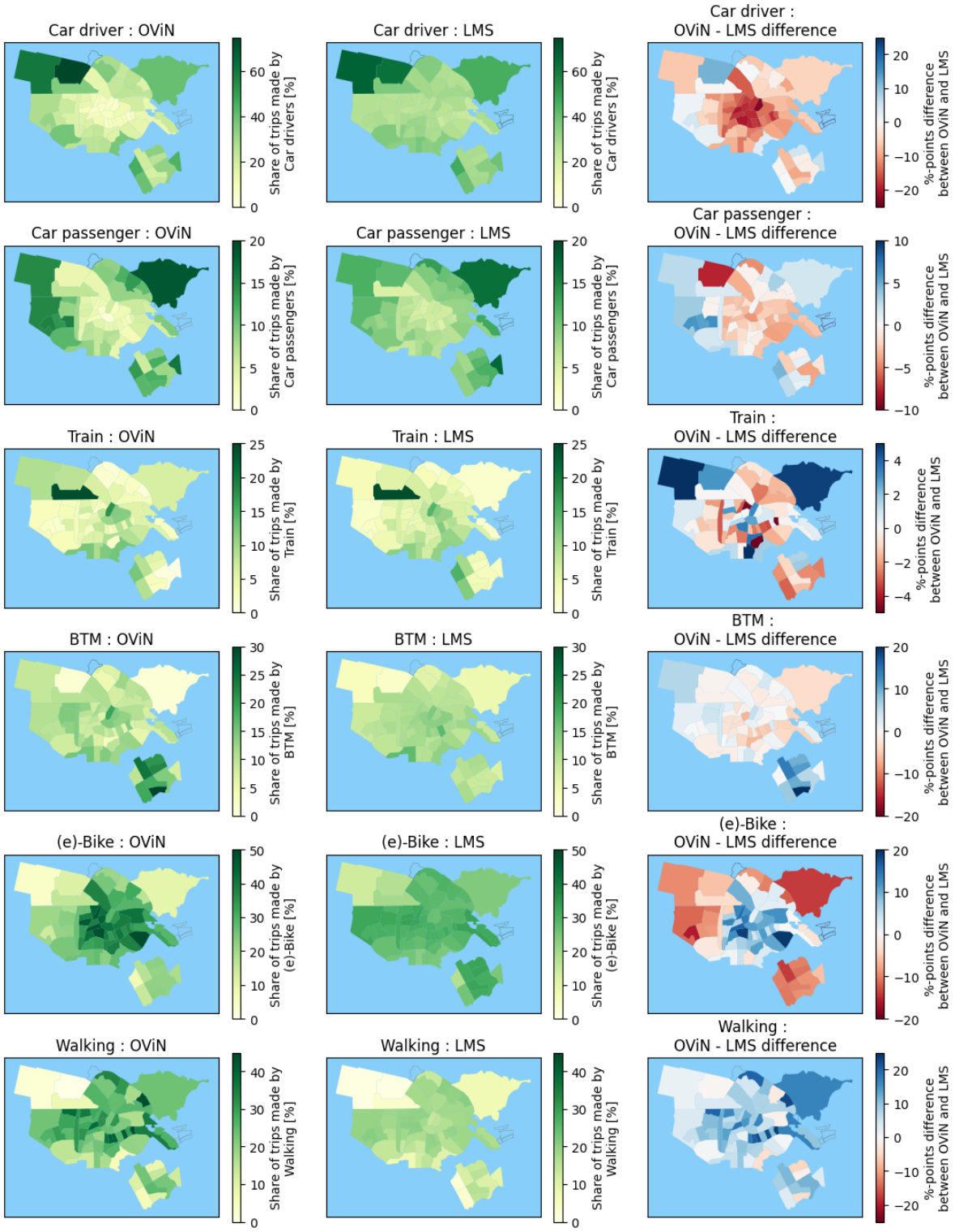

Figure 4.6: Left column: Modal split Amsterdam for OViN; Middle column: Modal split Amsterdam for LMS; Right column: Difference in modal split for LMS and OViN. All modal splits are based on the number of trips departing from a zone. Note: all difference plots use a different scale. This map is based on the combined OViN dataset for 2013-2017; the LMS OD-matrices (RWS WVL, 2018c) and (RWS WVL, 2020).

differences. It is important to note that because OViN is based on a sample, not all zones will be a good representation of the real modal split. So it could be more realistic to spread the travel behaviour a bit more over the different zones, as done in the LMS.

When looking at car travel, OViN shows a strong trend that car travel decreases the closer a zone is to the centre of the city. This effect however, is barely visible when looking at car travel according to the LMS, especially when looking at car drivers. This is shown in the difference plots in the right column, where the center shows a dark red color (so LMS overestimates car travel) and the red becomes lighter shade farther from the centre and even turns to blue (underestimation of car travel by the LMS). This implies that there are regional factors affecting car travel that are not yet included in the LMS. When looking at the DU of Amsterdam, a large part seems to have a DU of 6. However, car driver percentages according to OViN seem to differ from less than 10 % up to 40 % in zones with the same DU.

Public transport, in comparison, seems to be modelled more accurately. Especially when looking at the train, similar zones show high or low travel percentages of train travel, even though the absolute values are not always the same. There is no clear pattern visible on the map of which zones are over- or underestimated. BTM travel seems to be spread out a bit more, compared to OViN. LMS seems to strongly underestimate BTM travel in the south east of Amsterdam.

Bike use shows reverse trends compared to car use. The LMS strongly underestimates cycling close to the center of the city, while overestimating cycling farther away from the centre. Zones with a DU of 6 have shares of bike use between 20% and 50%, according to OViN.

Walking seems to be underestimated by the LMS in almost the whole municipality. There is not a very clear trend in the difference plot, as seen in the bike or car plots, though the predictions of the LMS seem to be a bit more accurate in the south and east of Amsterdam. Again, OViN shows large differences in travel behaviour in areas with the same DU.

To conclude, the LMS seems to model public transport the best and car drivers and cyclists the worst, when compared to OViN. There seems to be a clear pattern in the zones the LMS over- or underestimates. The LMS 'spreads' the travel behaviour out over the zones, which gives less differences between individual zones, when compared to the LMS. When comparing the travel behaviour with the DU, there seems to be a correlation (e.g. the center of Amsterdam has less car travel and more cycling, while the outer zones with a lower DU have higher car travel and less cycling), but the differences in travel behaviour between zones according to OViN cannot be captured by the DU alone. In other words, the DU does not show enough differentiation between the zones in Amsterdam.

4.2.3. Modal split The Hague, Zoetermeer, Leiden en Delft

Figure 4.7 shows the DU in The Hague, Leiden, Zoetermeer and Delft and figure 4.8 shows the modal split for those cities according to OViN, LMS and the difference between the two. The figure has a similar layout as the figure from Amsterdam. The difference shown here gives the absolute difference in mode share in %-points. For the relative differences, see appendix F.

The plot about the DU shows that The Hague has primarily a DU of 6, with a few zones with a DU of 5, while Leiden, Zoetermeer and Delft all have a DU of 4.

When looking at car travel, according to OViN, Zoetermeer clearly has higher levels of car driver use compared to (most zones of) the other cities. Car driver use in Leiden and Delft seem to be more similar to The Hague, even though they have the same DU as Zoetermeer. The LMS predicts similar levels of car driver use in all four cities, though car driver use in the Hague is slightly lower on average. It captures some trends seen in OViN (lower car use in the centres of The Hague and Leiden and high car in the bottom zones of Delft and Zoetermeer). The difference plot shows that the LMS highly overestimates car driver use in the centres of The Hague, Leiden and Delft, while underestimating most of Zoetermeer (although some centre zones are slightly overestimated). The Hague shows similar patterns as Amsterdam: car driver use in the centre is overestimated, while it is slightly underestimated farther from the centre.

The trends shown by car passenger use seem to be better captured by the LMS, though the absolute shares are still a bit off: the LMS seems to spread out travel behaviour again. Similar zones are over- or underestimated by the LMS, as observed for car driver use. Again, car passenger use for Leiden and Delft is more similar to The Hague than Zoetermeer, when looking at OViN.

For train use, the LMS models similar trends as seen in OViN. The Hague, Leiden and Delft all

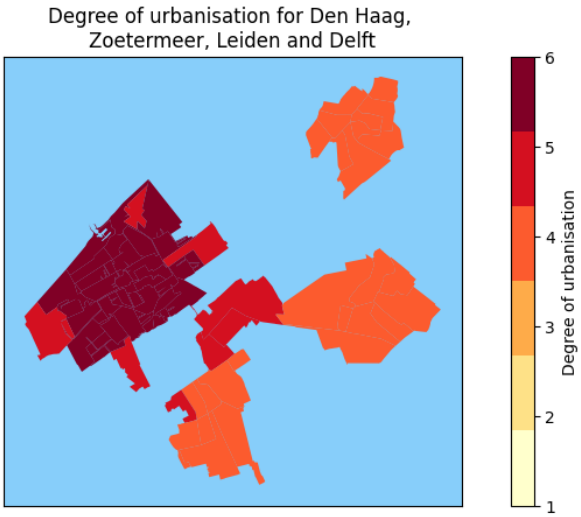

Figure 4.7: Degree of urbanisation in The Hague (top left); Leiden (top right); Zoetermeer (bottom right) and Delft (bottom left). This map is based on RWS WVL (2020).

show zones with high train use, while train use in Zoetermeer is low. This can probably be explained by the fact that Zoetermeer does not have an intercity train station, while the other cities do have one (or more).

When looking at BTM, OViN data shows that BTM use in the south east of the Hague, is more than double, compared to the other cities. Where Zoetermeer shows a share of BTM higher than 5%, Leiden and Delft have BTM use mostly under 5%. Interestingly, LMS models similar levels of BTM use in The Hague, Delft and Zoetermeer, even though the share of BTM in Delft is more similar to Leiden and The Hague has a way larger share of BTM, according to OViN. A possible explanation for this could be that the LMS pays more attention to the availability of BTM and less to the quality of the BTM. To illustrate, The Hague has several tram lines and a metro line. Zoetermeer has a metro line and Delft a tram line. Leiden only has busses. Based on this information alone, the LMS predictions appear more logical. However, there might be large differences in the quality of the different BTM networks and the attractiveness of other modes. When looking at the differences between OViN and LMS, it shows that BTM is overestimated almost everywhere, except for the centre of The Hague. Especially in Delft and Zoetermeer the use of BTM is overestimated.

For cycling, the reverse is seen compared to share of car driver. OViN shows clear areas with very high bicycle use (40+% of the trips) and areas with very low use (only 10%). The LMS however, predicts similar levels of bike use in all four cities. Because of this, bike use is overestimated in Zoetermeer and the centre of The Hague, while being underestimated in Delft and Leiden. The overestimation of bike use in the centre of The Hague is an interesting difference with Amsterdam, where cycle use was underestimated. When comparing those two cities, both The Hague and Amsterdam seem to have a different modal split, even though the DUs are the same.

Finally, the share of walking. In all four cities, similar levels of walking can be observed, with a large peak around the centre of The Hague. The LMS also predicts similar shares of walking for all cities, though the peak, as seen in the Hague, is not modelled. Overall, walking is underestimated almost everywhere. This is in line with the earlier observations in Amsterdam and the modal split per DU for the whole country.

To conclude, even though Leiden, Zoetermeer and Delft have the same DU, travel behaviour in Zoetermeer seems to be very different compared to Delft and Leiden. The latter two are more similar to The Hague, except when looking at BTM. The LMS, however, often predicts similar mode shares for all four cities, except for public transport. All in all, it seems that there are differences in the spatial environment between those cities, that are currently not captured in the LMS.

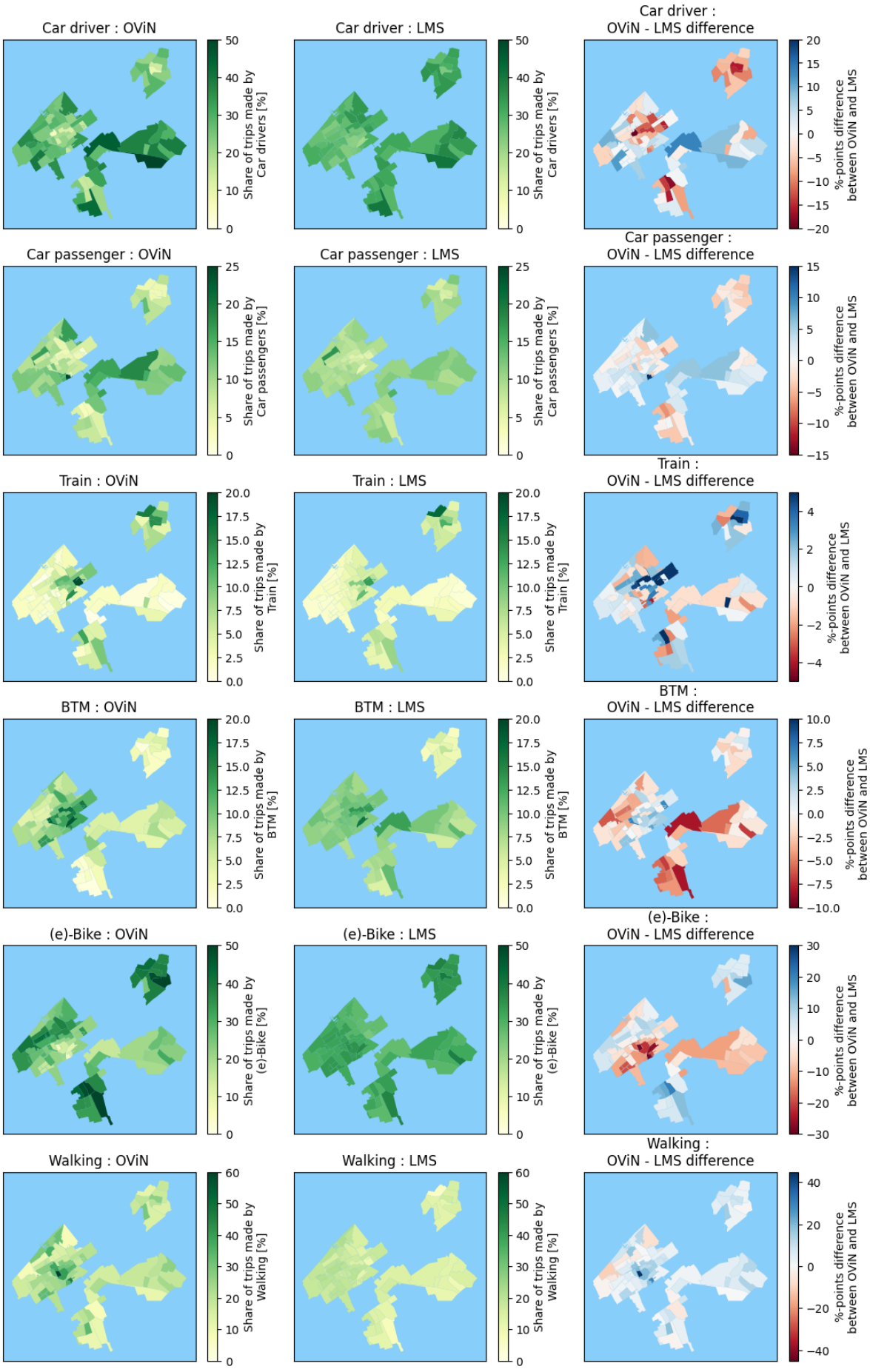

Figure 4.8: Left column: Modal split The Hague (top left); Leiden (top right); Zoetermeer (bottom right) and Delft (bottom left) for OViN; Middle column: Modal split The Hague, Leiden, Zoetermeer and Delft for LMS; Right column: Difference in modal split for LMS and OViN. Note: all difference plots use a different scale. This map is based on the combined OViN dataset for 2013-2017; the LMS OD-matrices (RWS WVL, 2018c) and (RWS WVL, 2020).

4.2.4. Modal split Zeeland

As a final example, an area outside the Randstad is analysed. Figure 4.9 shows the DU in the province of Zeeland and figure 4.10 shows the modal split for Zeeland according to OViN, LMS and the difference between the two. The figure has a similar layout as the figure from Amsterdam. The difference shown here gives the absolute difference in mode share in %-points. For the relative differences, see appendix F.

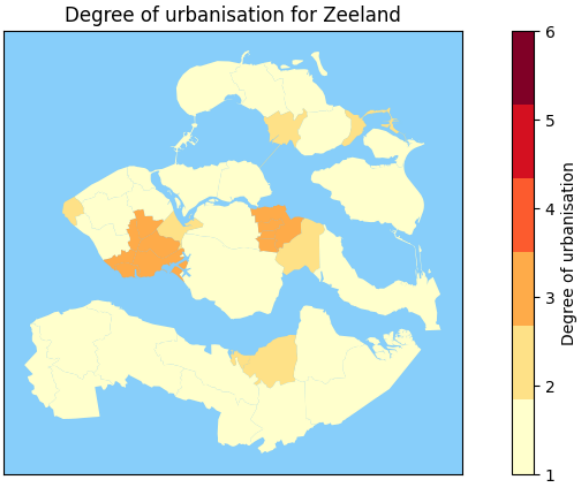

Figure 4.9: Degree of urbanisation in Zeeland. This map is based on RWS WVL (2020).

Zeeland has a DU of 1 for most zones, with some zones having a DU of 2 or 3. The zones with a DU of 3 are the cities Middelburg and Goes. Compared to the zones in the previous example, the zones in Zeeland cover a lot more area. There are of course also less people living in the same area[4].

Similarly to the examples of the Hague and Amsterdam, in general the LMS seems to 'spread out' travel behaviour more, while OViN shows more highs and lows. In general car use (both car driver and passenger) is very high. The southern part shows the highest car driver use, while the middle peninsula has the lowest car use, especially around the larger cities. Still, the car use is a lot higher than in the cities in the Randstad. The LMS seems to overestimate car travel of the middle peninsula, while underestimating the rest.

Train and BTM seems to be modelled well by the LMS in general. A few zones in OViN seem to be outliers, but this seems to be corrected by the LMS. This is an advantage of the tendency of the LMS to 'spread out' travel behaviour. Walking is again mostly underestimated by the LMS, though the LMS does seem to follow some trends (e.g. higher walk share around the cities).

Cycling again shows large errors. OViN shows that there is a of difference in bike use (a share of around 50% around Middelburg and surroundings and a share close to 10% in the southern part of Zeeland). The LMS however, predicts similar levels of bike use, overestimating some zones, while underestimating others.

To conclude, similar to the previous analyses, bike and car driver use seem to be the two modes that show the largest differences between OViN and LMS in Zeeland. This implies that these two modes could use the most improvements. The differences in modal split as seen by OViN can not be captured by the DU alone.

[4] This does not mean that the data from Zeeland is less reliable. The modal split of both the municipality of Amsterdam and the province of Zeeland are based on approximately 17,000 OViN trips, while they consist of 68 and 42 zones respectively. This means that in Zeeland, more trips are gathered per zone on average.

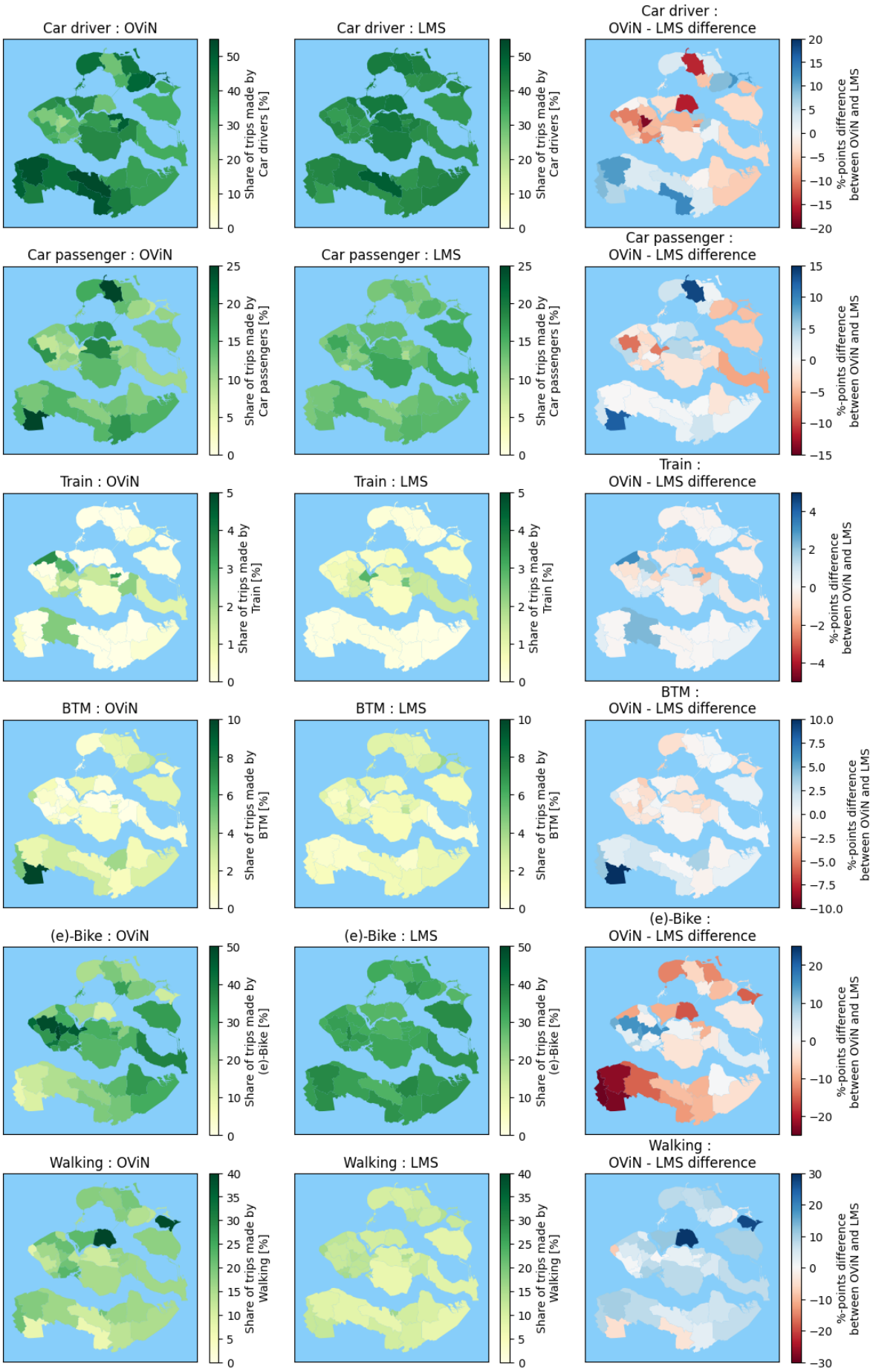

Figure 4.10: Left column: Modal split Zeeland for OViN; Middle column: Modal split Zeeland for LMS; Right column: Difference in modal split for LMS and OViN. Note: all difference plots use a different scale. This map is based on the combined OViN dataset for 2013-2017; the LMS OD-matrices (RWS WVL, 2018c) and (RWS WVL, 2020).

4.2.5. Conclusions exploratory analysis modal split

This subsection will summarize the most important findings from the exploratory analysis.

- Both on national and on regional level, the LMS seems to overestimate car driver use and underestimate walking.
- The average bike use seems to be modelled relatively well on national level when looking at the DUs and does not vary much. However, according to OViN bike use can vary a lot regionally. This is not modelled accurately by the LMS. This trend can be seen in large cities (Amsterdam, The Hague, Leiden, etc.), but also in more rural areas (Zeeland)
- Train use is modelled relatively well. Though the absolute share might be off, zones with high or low use are often identified. This is also partly true for BTM, though the different trends in BTM seem to be captured worse than train use (e.g. see South East Amsterdam or Delft and Zoetermeer). A possible explanation for this could be that there are more variables included in the LMS that are related to train use, compared to BTM.
- The LMS seems to 'spread out' travel behaviour more, predicting similar levels of mode use in neighbouring areas. This can be a good thing (e.g. removing outliers due to the lack of reliable OViN data of a zone), but interesting trends are lost that can be seen in the OViN data (e.g. car driver and bike use in Amsterdam that change a lot, the farther from the city centre). There are often large areas with the same DU. If there are not enough other variables to identify differences between those zones, the LMS will predict similar travel behaviour.
- Places with the same DU do not necessarily display the same travel behaviour (e.g. Leiden, Zoetermeer and Delft). This last observation might be the most important one and strengthens the hypothesis that using only the DU in transport models is insufficient to capture the differences in travel behaviour between different regions.

These findings will be important for answering the third and fourth sub-question and at the end the main research question. Besides that, the findings from this section will provide additional criteria that can be used in the cluster analysis. The differences in modal split observed between Delft, Leiden and Zoetermeer indicate that those three cities should not belong to one cluster. See section 4.3.1 for a further explanation of the cluster criteria. The differences in modal split that were found between Delft, Leiden and Zoetermeer will form one of the criteria that are used in the cluster analysis, next section.

4.2.6. Other aspects of travel behaviour

The analysis in this thesis focuses primarily on the modal split. An attempt was made to analyse other aspects of travel behaviour (i.e. travel distance, travel time and part of the day). Due to limitations in the scope, this was not further elaborated. See the discussion (section 5.1) for more details about this process and appendix G for some initial graphs that were made.

4.3. Cluster analysis results

The goal of the cluster analysis is to create regions by clustering zones based on D-variables that show significant differences in travel behaviour. These regions should be better at differentiating travel behaviour than the DU or should identify regions with interesting travel behaviour, that cannot be captured with the DU alone. To make the clusters, hierarchical clustering will be used. This section will first give a description of the process that was used to determine the optimal clusters. After that, those best clusters will be analysed. In the next section, the propensity score matching is performed and the different clusters are analysed.

4.3.1. Clustering process

This subsection gives a description of the cluster process and the choices and assumptions that had to be made.

Preparing the D-variables

Before the start of the clustering, the D-variables need to be prepared. The gathering process of those variables is described in section 4.1.2. However, these data contained some missing values. For example, the CBS does not publish personal data due to privacy reasons if it has less than 5 observations (Van Leeuwen & Venema, 2023). It is not possible to perform the clustering with any missing data. 12 of the 48 variables in total contained some missing values. Because the number of missing values for those variables was low (0.5 - 2% of the zones), it was decided to use average values for those variables as a replacement. This way, the zones with missing values could still included in the clustering process. Because these zones often had a low total population (otherwise data like the number of houses with a certain built year would not have been censored), their impact on the average travel behaviour in a cluster would be low.

The next step is to scale the values from each variables to the range of 0 to 1. This is done to avoid that a variable with very high values (e.g. the population density in persons / ha) has a larger impact on the clustering process, compared to a variable with very low values (e.g. the entropy index which ranges from 0 to 1) (de Souto et al., 2008).

After that, the variables are ready to be clustered.

Analysis of D-variables

Because there is not one easy indicator to optimize during the clustering process, finding the best clusters will be a manual process. To make this process easier, correlation matrices were made of the correlation between each variable and the correlation of each variables with the modal split, using the Pearson correlation coefficient. The Pearson correlation coefficient gives a score ranging from -1 to +1. If two variables have a high score (near +1) the two variables are positively linearly correlated. A low score (near -1) implies a negative linear correlation and a score near 0 means that there is little or no linear correlation. Equation 4.6 shows how to calculate the Pearson correlation coefficient (Berman, 2016). Here x and y are the D-variables that are being compared, i is the LMS zone ($i = 1, 2, ..., I$) and $\overline{x}$ and $\overline{y}$ are the average values of those D-variables.

$$\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2}\sqrt{\sum (y_i - \overline{y})^2}} \tag{4.6}$$

The black squares in the correlation matrices in this thesis imply that, based on the data, we cannot say if the observed correlation is due to chance or due to a relationship between the variables (two-tailed p-value > 0.05) (James et al., 2023, p. 77).

Figure 4.11 shows the correlation of all variables with each other. Blue means a positive correlation, red a negative correlation and black that it is unknown if there is a relationship between the variables.

Several things can be noticed while looking at the matrix. First of all, several ‘blocks’ of positive correlation can be seen on the diagonal. These are often a set of variables belonging to one D-variable that are highly correlated with each other (Density, Destination accessibility and Distance to transit). This implies that it might not be needed to include each individual variable from that D-variable (e.g. maybe the variable ‘distance to point of interest’ captures the D-variable Destination accessibility good enough and additional variables like ‘distance to recreation’ are not needed).

Secondly, the Density variables seem to be highly (positively or negatively) correlated with a lot of the other variables. Based on this, it can be argued that Density is indeed a good variable to include in transport models to capture the spatial environment. For example, the variable ‘distance to point of interest’ might provide additional information about the zones, which can make a transport model more accurate. However when lacking data about distances, a Density variable might be an acceptable proxy due to the strong (negative) correlation.

In other words, when including only Density variables a lot of the other D-variables are also implicitly captured to a certain extent. This can be a good reason to implement a variable like the DU.

Figure 4.12 shows the correlation matrix of each variable with the modal split. Again several things can be noticed by looking at this matrix.

First of all, the Density variables seem to be highly correlated with travel behaviour. This strengthens the point made in the previous paragraph that Density variables are suitable to use in transport models

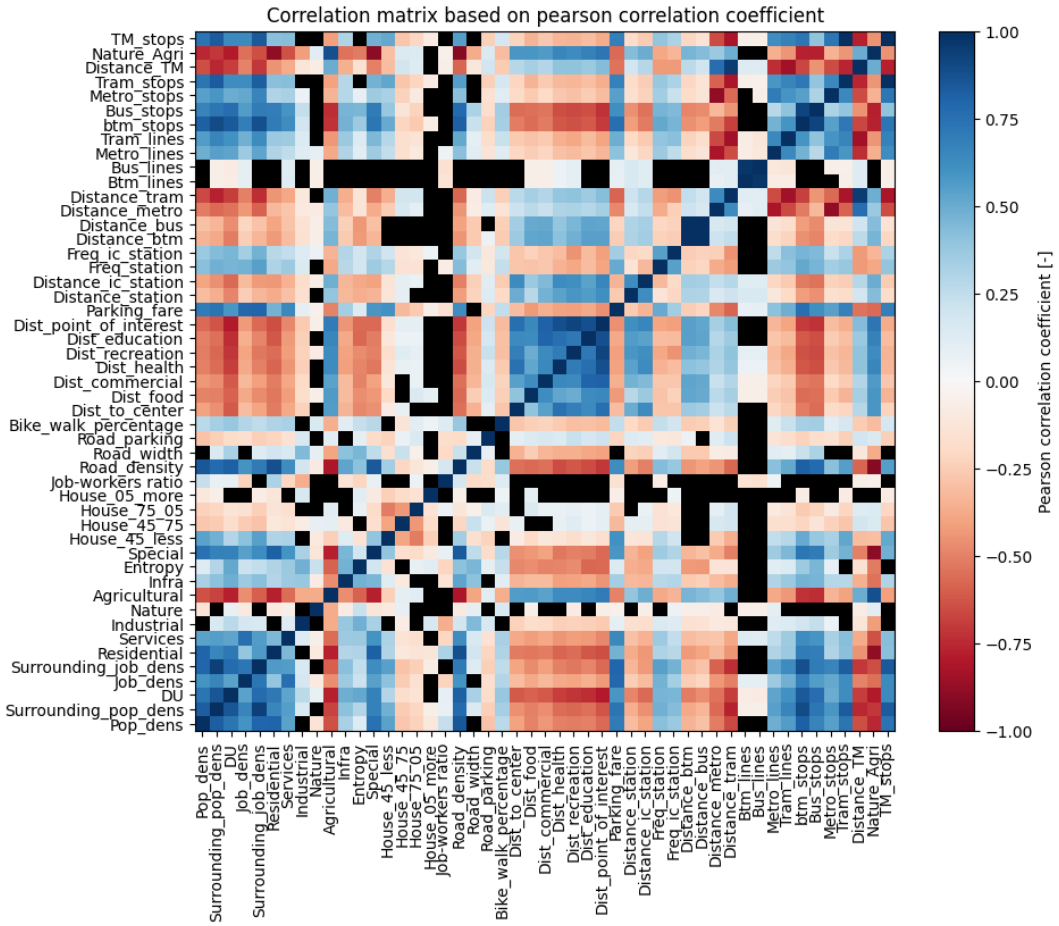

Figure 4.11: Correlation matrix for all variables, based on the Pearson correlation coefficient. For the description and source of each variable, see table 4.1.

and can serve as an acceptable proxy when other data is not available. They seem to have the strongest correlation with car driver trips and BTM trips and the weakest correlation with bike trips. Even though all Density variables show a similar trend, differences can still be seen. For example, while job density has a weaker correlation with most modes, their correlation with train use is the strongest from all Density variables. This implies that it can be valuable to include more than 1 Density variable (e.g. both job density and population density), because they can still capture slightly different effects.

Secondly, most variables seem to show a similar trend: they are positively correlated with car use and negatively correlated with all other modes, or the reverse. This result is in line with expectations from the literature review: dense areas with more public transport options and shorter distances to activities, stimulate higher use of public transport, walking and cycling and are less attractive for cars. The strengths of the correlation however, show more variation, which can still make it valuable to use multiple variables in a transport model. There are a few variables that show a different trend (e.g. industrial land use), though those correlations are often very weak.

Bike use seems to be the mode that has the weakest correlation with the different variables. This observation is in line with figure 4.4. This figure shows that according to both LMS and OViN, bike use seems to vary little over the different DUs and is relatively high (30%) throughout the whole country.

Get familiar with different clusters

Before the ‘final’ clusters are made, it is important to do some tests. This is done to test if all the code works as intended and to get a first idea of the effect of the different D-variables.

Different sets of variables are clustered to see if interesting patterns appear. These variables are chosen semi-randomly (i.e. some variables are tested that seemed promising according to the litera-

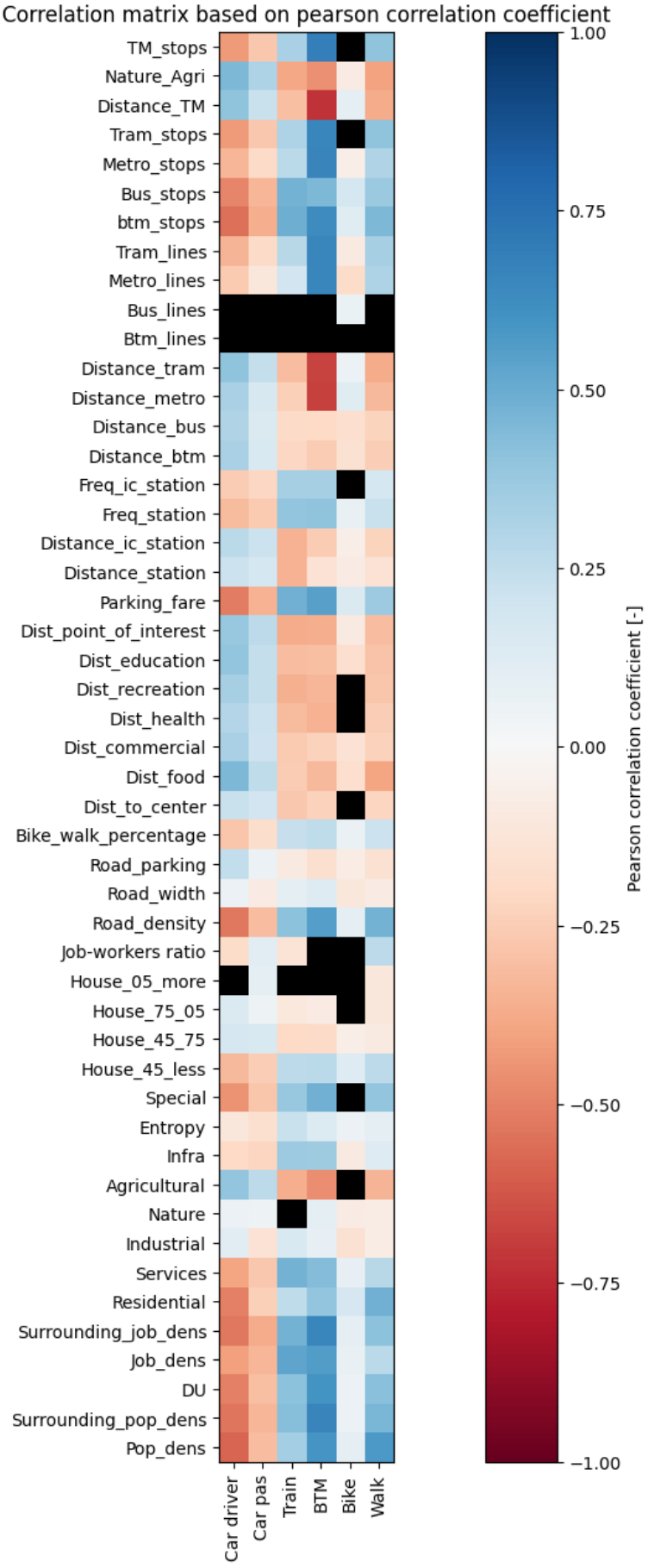

Figure 4.12: Correlation matrix of all variables with the modal split, based on the Pearson correlation coefficient. For the description and source of each variable, see table 4.1. The modal split is based on the combined OViN dataset for 2013-2017.

ture, like entropy, and some variables from all 6 D-variables are included).

By looking at those different test cluster sets, two interesting types of clusters are discovered: the 'medium-sized cities' and the 'suburbs'. The testing shows that those medium-sized cities often have a DU of 4 or 5 and have share of bike use close to 35%, compared to the 30 % of the other clusters. This is an unusual discovery, because the correlation between bike use and the different D-variables is relatively low. The LMS highly overestimates car driver use in these zones (around 10%-points) and predicts it to be higher than in the suburbs. However, according to OViN car driver use in the medium-sized is lower than in the suburbs. The suburbs are the zones around the very large cities (DU of 6) that show high BTM use, low bike use and fairly low car use. Those suburbs also have a DU of 4 or 5.

To summarize, the medium-sized cities and the suburbs discovered during the cluster testing, show interesting travel behaviour, that is not always captured by the LMS. Both clusters have a similar population density, but still show very different travel behaviour. They do not seem to follow the general trend as the DUs (i.e. a decrease in car travel means an increase in public transport and walking, while bike use barely changes). These clusters can be further optimized for the final cluster sets.

Cluster criteria

Section 2.5.3 gave an overview of criteria and indicators to use in the clustering process. To summarize, use the correlation between the variables and the correlation of the variables with travel behaviour; evaluate the distribution of the variables within a cluster; look at variance in travel behaviour; a cluster should not contain too little zones; the numbers of clusters should not be too high or too low.

Based on the insights obtained from the data analysis up to this point, additional and more specific criteria can be added:

- Section 4.2.3 found that Delft and Leiden show similar travel behaviour according to OViN, while Zoetermeer shows different trends. One of the indicators to check the quality of the clusters is to see if (the centers of) Delft and Leiden belong to the same cluster, while Zoetermeer belongs to a different cluster. This was difficult to do during the initial tests. Delft and Zoetermeer often belonged to the same cluster, while Leiden belonged to a different one. This was presumably due to the fact that Zoetermeer and Delft both have a tram or metro line, which got them in a 'suburb' cluster. However, according to figure 4.8, BTM use in Delft is lower than in Zoetermeer and more similar to Leiden when looking at OViN. According to the LMS, BTM use in Delft en Zoetermeer is on a similar level and higher than in Leiden. A possible cause of this is that the LMS overestimates BTM use in Delft due to the presence of a tram line.

  To conclude, the way those three cities are clustered, will be a good indicator to judge the cluster sets.

- The insights obtained during the cluster testing can also be used as a criteria. The final clusters should make a clear distinction between the suburbs and the medium-sized cities.

- When testing different clusters, some clusters showed high errors in the LMS data and did not follow the general trend of the OViN data (as seen with the suburbs and medium-sized cities). Those errors imply that the LMS is not good in capturing average travel behaviour in that combination of zones, which makes those clusters extra interesting. If there is a chance in making clusters that can improve the LMS, it is important to find clusters that show errors. This can be another indicator of interesting clusters.

- During the cluster testing, it often happened that (large) cities got a lot of different clusters, while more rural areas all belonged to the same cluster. It is important to find a good balance between keeping enough detail in the highly urban areas, while not under fitting the rural areas.

Finding the final clusters

Because there are no clear indicators that can be maximized (or minimized), finding the optimal number of clusters and variables to include is a manual process. Each cluster set is evaluated by the author, based on the criteria that are developed. The remaining of this subsection gives a general description of this process and its challenges.

First, all variables are sorted based on the strength of their (average) correlation with travel behaviour. One by one, new variables were added in order of correlation. When a variable seems to

have a positive effect on the clusters, based on the different criteria, the variable is kept. Otherwise it is removed. The real process turned out to be less linear than simply adding variables in order of correlation. Sometimes already discarded variables are brought back to see their effect in a new cluster combination. In other cases different weights are given to variables (this will be elaborated later). Besides looking only at the average correlation of each variable with travel behaviour, variables are also sorted on their correlation with a single mode. This helps with choosing new variables to add (e.g. a set of clusters gives good result, but the medium-sized cities seem to disappear a bit. By adding a variable that shows a (relatively) high correlation with bike use, it was possible to put more emphasis on those medium-sized cities, because they also show a lot of bike use.)

During the process, it is discovered that some variables do not produce the desired effect. These variables are improved. For example, there are variables for the distance to various BTM stops and the number of different directions for each stop. However, this variable gives too little information about the quality of the BTM and zones close to a tram or metro stop are automatically placed in separate clusters from the rest of the Netherlands. This does not seem realistic and these variable are replaced with a new variable that gives the number of BTM stops within a certain radius of the zone centroid. These variables give a better indication of both the quality of the BTM network (more stops means a more advanced transit network) and the accessibility (more stops means shorter access and egress time on average). Adding these variables improves the clusters and brings more nuance (e.g. Delft stops being in a 'suburb' cluster).

For each set of variables, different numbers of clusters are tested to see how this affects the clusters. The different indices (silhouette score; Calinski-Harabasz score; Davies-Bouldin score) favoured almost always 2 or 3 clusters. These indices favour clusters that are very distinct. However, the goal of this analysis is not to find clusters of which the spatial environment is as different as possible, but to find clusters that show interesting differences in travel behaviour and are an improvement on the DU. In other words, it does not matter if the spatial environment of two clusters is similar in some ways, as long as they display significant differences in travel behaviour. Only 2 or 3 clusters is too few for the purpose of this study, because they are not an improvement on the 6 DUs. The indices were deemed unsuitable and were mostly ignored. The ideal number of clusters for each set of variables was based on what the author thought provided a good balance between enough distinction between regions (both urban and rural clusters had to be represented), while not overfitting. This often resulted in 6 or 7 clusters.

Two different sets of clusters are made. The first is as simple as possible: this set of clusters is based on as few variables as possible to make them simple to create and implement them in a transport model. It is interesting to see if, by adding only a few other variables besides population density, the clusters are better in capturing travel behaviour than the DU. This cluster set is known as the unweighted cluster set

The second set of clusters is more complicated and has the goal to capture more different aspects of the spatial environment and to capture more subtle differences between zones. They are made with the assumption that each D-variable is equally important in distinguishing zones and has an equal weight in the clustering process. However, each D-variable can be made up of several variables. For example, the Density variable consists of both population density and job density, that both have a weight of 1. The Design variable consist of only the road density, with a weight of 2. This makes both D-variables equally important. This cluster set is known as the weighted cluster set.

The list of all possible combination of variables that are tried can be found in appendix I. At the end, two sets of clusters are found that were in line with the criteria. These will be further analysed in the next section.

4.3.2. Analysis of cluster sets

This section will first give an analysis of the weighted cluster set and after that an analysis of the unweighted cluster set. At the end, these two cluster sets are compared with each other and with the DU.

Analysis of weighted cluster set

The weighted cluster set was made by clustering the variables shown in table 4.3. All 6 D-variables are represented in this cluster set. This was not a requirement for the clusters, but each D-variable seemed

Table 4.3: Overview of different variables, their D-variables and the variable weights that were used to create the weighted cluster set. The sources of each variable can be found in table 4.1.

D-variable	Variable	Weight
Density	Population density	1
	Population density, including surrounding zones	1
	Job density	1
	Job density, including surrounding zones	1
Diversity	Share of service land use	2
	Houses built before 1945 ratio	2
Design	Road density	4
Destination accessibility	Distance to point of interest	4
Distance to transit	Number of bus stops	2
	Number of tram/metro stops	2
Demand management	Parking fare	4

Table 4.4: An overview of the 7 clusters with the names and the size of the clusters from the weighted cluster set.

Cluster number	Cluster name	Number of zones	Share of zones
1	Centres of large urban areas	73	5.2 %
2	Centres of medium-sized cities	85	6.0 %
5	Suburbs of large urban areas	59	4.2 %
6	Older towns/ suburbs	90	6.4 %
3	Suburbs of medium-sized cities	312	22.2 %
4	Towns & small cities	280	19.9 %
0	Rural areas	507	36.1 %
Total		**1406**	**100 %**

to add something valuable. The choices of the different variables will be further elaborated below. For this cluster set, 7 clusters seemed to give the best results, contrary to the results from the different indices (e.g. silhouette). See appendix H for the values of the different indices and the dendrogram. Table 4.4 shows an overview of the size and name of each cluster. Figure 4.13 shows the different clusters on a map of the Netherlands and figure 4.14 shows how the different variables are distributed in each cluster. Figure 4.15 shows the differences between this cluster set and the DU and figure 4.16 shows the modal split for the clusters. The clusters are sorted based on the car driver use according to OViN (see figure 4.16).

Next, an analysis is done for each cluster, starting with the cluster with the lowest car use.

- **Cluster 1: Centres of large urban areas**

  The zones in this cluster form the centers of the three largest cities: Amsterdam, The Hague and Rotterdam. This cluster is made up almost fully of zones with a DU of 6 and is characterized by a very high density (both job and population), a high parking fare and a high road density. There is a large tram and/or metro network, which explains the high BTM use. There are also many bus stops, though less than in the centres of medium-sized cities. The share of land use used for services (e.g. shops, restaurants, libraries) is high and similar to the centres of medium-sized cities. The ratio of houses built before 1945 is also high and the average distance to points of interest is very low. None of these variables seem out of place.

  When looking at the modal split, it shows that car driver use is extremely low with only 16% of the trips on average. However, it is overestimated by the LMS, with more than 70%. This is in line with earlier observations in section 4.2.1, where car use in the city centre of Amsterdam was overestimated. Car passenger use is also low, but only slightly overestimated by the LMS. Both train and BTM use are very high compared to the other clusters and both the absolute and relative

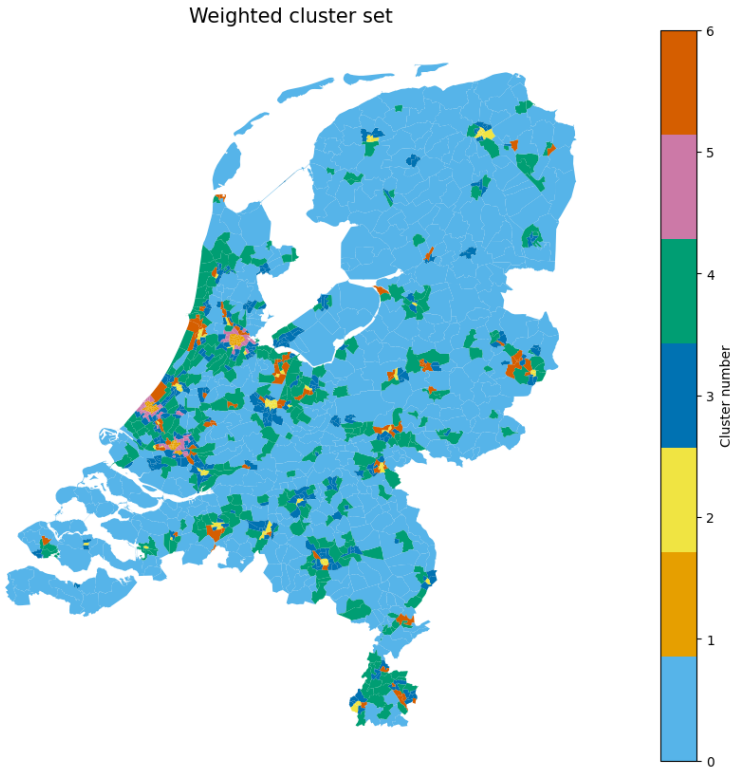

Figure 4.13: Map of the Netherlands displaying the weighted cluster set. This map is created using hierarchical clustering. For the variables used to create this cluster set see table and 4.3. The shapes of the zones are based on RWS WVL (2020).

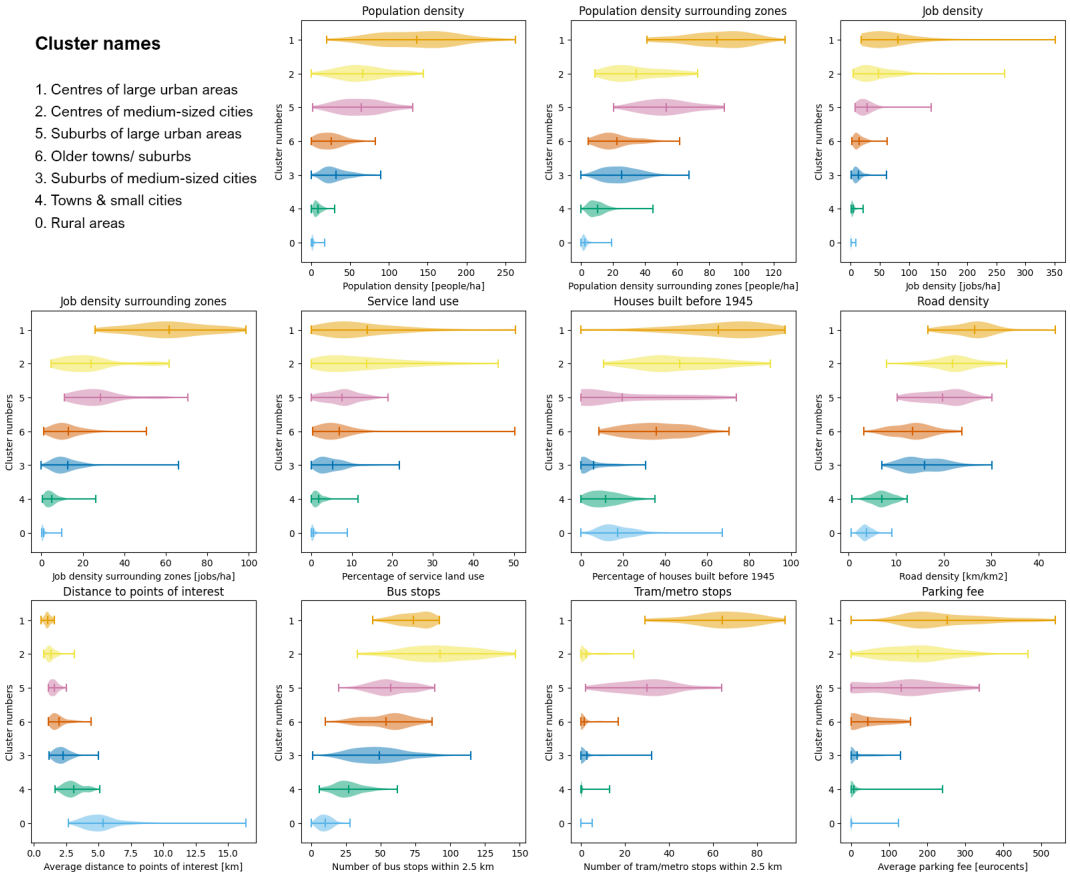

Figure 4.14: Distribution of the variables in each cluster for the weighted cluster set in a violin plot. This figure is based on RWS WVL (2020) and the variables from table 4.3.

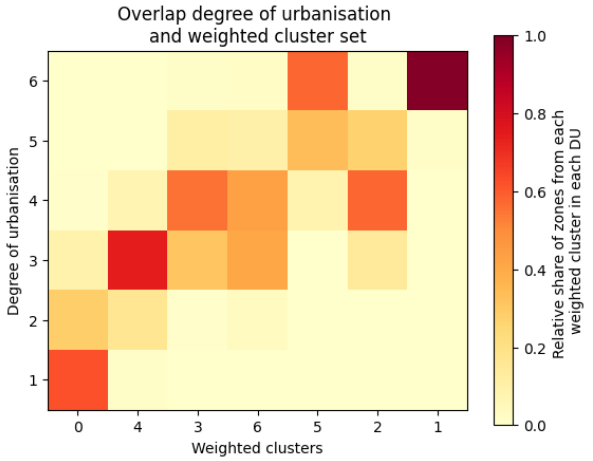

Figure 4.15: Comparison of the degree of urbanisation with the weighted cluster set. Each square shows the share of zones of a certain cluster, belonging to a certain DU. For example, of the zones in cluster 2, around 60% has a DU of 4, around 25% has a DU of 5 and around 15% a DU of 3. This figure is based on RWS WVL (2020) and the variables from table 4.3.

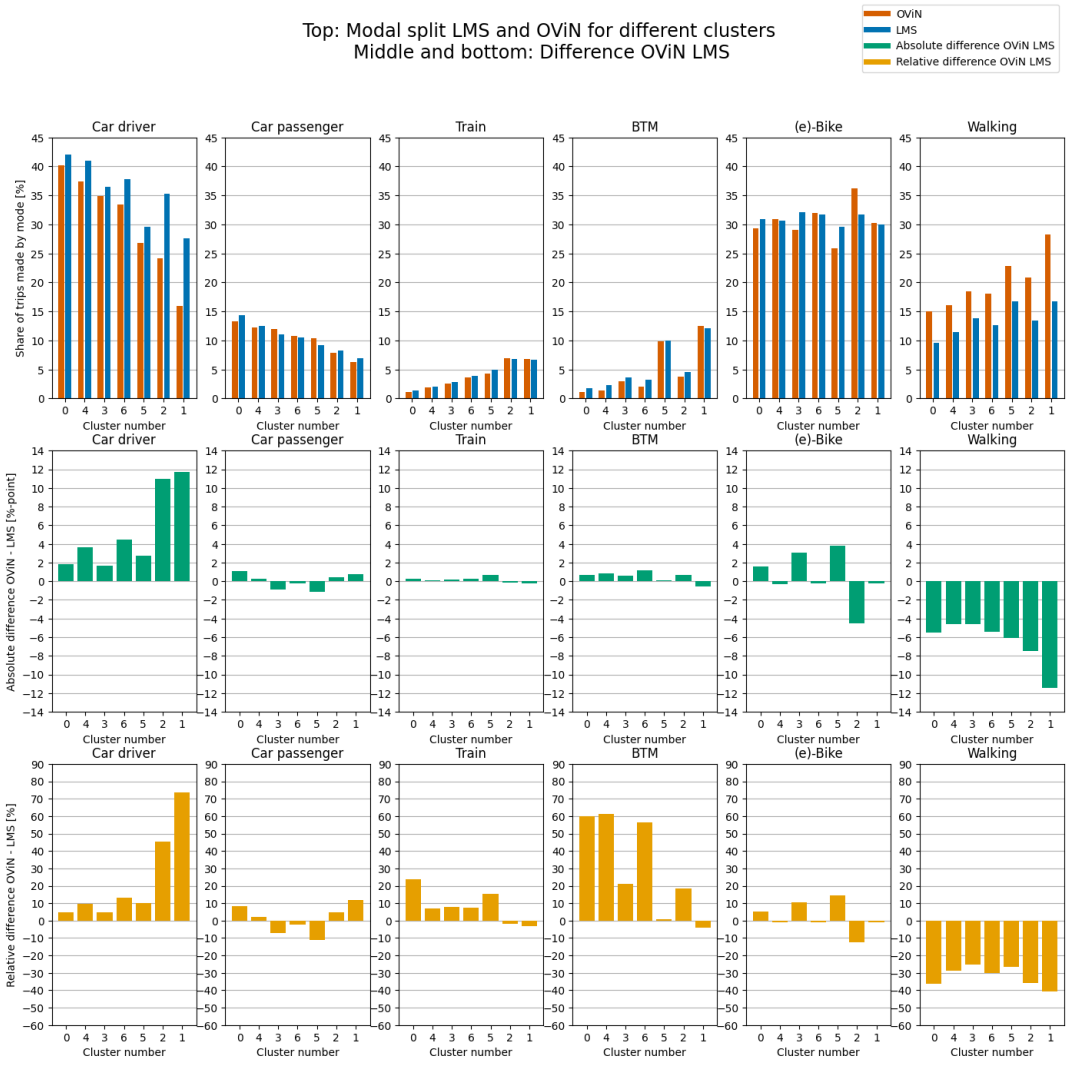

Figure 4.16: Top: Comparison between the modal split for different clusters of the weighted cluster set for OViN (red) and LMS (blue). The percentages are based on the number of trips. Middle: Difference in %-points between OViN and LMS (OViN minus LMS). Bottom: Relative difference between OViN and LMS (difference = (LMS - OViN) / OViN). This graph is based on the combined OViN dataset from 2013-2017, the LMS OD-matrices (RWS WVL, 2018c) and clusters from the weighted cluster set, which are based on the variables in table 4.3.

difference between OViN and the LMS predictions are low. Bike use seems to be very average, compared to other clusters and also predicted accurately. The share of walking in this cluster is the highest of all clusters, though the LMS predicts it to be almost 12 %-points lower.

- **Cluster 2: Centres of medium-sized cities**

  The zones in this cluster mostly form the city centres of medium sized cities like Leiden, Zwolle, Arnhem or Groningen. They primarily consist of only a few zones together at the same location. These clusters are characterized by a relatively high population density, similar to cluster 5, which can be characterized as the 'suburbs of large urban areas'. The population density when including the surrounding zones, however, is significantly lower than cluster 5. A similar effect can be seen when looking at the job density. This can be explained by the smaller size of these cities, compared to large urban areas in cluster 1 and 5. Medium-sized cities are often surrounded by more rural areas and lower density 'suburbs'. The cluster is mostly made up of zones with a DU of 3, 4 or 5.

  When looking at averages of variables like parking fee, road density, houses built before 1945, and distance to point of interest, this cluster fits nicely in the gradual decreasing (or increasing in the last case) trend starting from cluster 1, the centres of large urban areas. There are, however, some irregularities. The number of tram/ metro stops seems to be close to 0, while the average number of bus stops is the highest from all clusters. The share of land used for services is high and on a similar level as cluster 1.

  As identified in the testing phase, medium-sized cities show some interesting travel behaviour. Car driver use is relatively low at only 24% though it is estimated by the LMS to be 35%. The LMS predicts that car driver use is almost 6 %-points higher than in cluster 5, though OViN shows that the share of car drivers is lower than in cluster 5. A possible cause for this prediction is the low surrounding population density of the medium-sized cities, which gives a lower DU. Train use however is high at 7% and is accurately predicted by the LMS. A possible cause of the high train use is that most medium-sized cities have an intercity station, which due to the smaller size of the city is close and easy to reach. BTM use is around 4%, which is very low compared to cluster 1 and 5. This is because of the lack of high quality transit (tram/metro).

  Bike use shows the most interesting results, due to the highest share of all clusters. Because the LMS seems to predict similar levels of bike use in all clusters, bike use in this cluster is underestimated with 5 %-points. Looking back at the study of the LMS documentation in section 3.3, it was found that of the factors related to the spatial environment, relatively few are related to bike use. What makes this high use in bike extra interesting is that bike use seemed to be largely unaffected by the spatial environment, based on the earlier analysis for the modal split for each DU and the correlation matrix for the variables with the modal split. Even though bike use does not show strong trends in most of the country, this cluster seems to be an exception.

  Walking in this cluster seems to be higher than average, though a bit lower than in clusters 1 and 5. It is also underestimated by the LMS, though the same trends (a slight dip for walking in this cluster, compared to cluster 1 and 5) is modelled by the LMS.

- **Cluster 5: Suburbs of large urban areas**

  The zones in this cluster are the neighbourhoods around cluster 1, the city centres of Amsterdam, The Hague and Rotterdam. The zones mostly have a DU of 5 or 6. The cluster is characterized by high densities and a surrounding zones job and population density that is higher than in cluster 2, the medium-sized city centres. This indicates that cluster 5 belongs to larger high density areas. The (average) parking fee, road density, share of houses built before 1945 and the distance to points of interest does not show irregularities and still decreases (or increases) almost linearly, starting from cluster 1. While clusters 1 and 2 are, for a big part, made up of older city centres, cluster 5 seems to be made up of newer neighbourhoods made for the growing cities. This gives a low share of old houses. Interestingly, while the average distance to points of interest is strongly correlated with the DU and other Density variables (see figure 4.11), the average distance to points of interest for cluster 5 is larger than in cluster 2. In other words, a higher Density is not directly equivalent to a shorter distance to points of interest.

The number of tram and metro stops is very high, indicating a complex tram and metro network, although there are less stops compared to cluster 1. The number of bus stops is also high, though lower than cluster 1 and 2. The lower number of bus stops than in cluster 2 can be explained by the higher density tram/metro network. The land used for services shows a decrease compared to cluster 1 and 2 on average, indicating that there are less shops and similar buildings in this cluster. However, the high share of service land use in clusters 1 and 2 can partly by explained by some outlier zones, so the medians of the clusters 1, 2 and 5 are closer together.

The modal split in this cluster seems to mimic travel behaviour in cluster 1 to a certain extent. Car use, for both drivers and passengers is higher than in cluster 1, while the use of all other modes is lower. The share of bike trips in this cluster is the lowest of all clusters. When comparing bike and BTM use for the clusters 2 and 5, it implies that BTM and bike compete with each other to a certain extent. However, when comparing cluster 5 with the other, more car-centered, clusters (0, 4, 3, 6) the increase in bike use in those clusters is smaller than the decrease in public transport use. This implies that BTM not only competes with bike use, but also with car use (which is 5 to 15 % points higher in the clusters 0, 4, 3 and 6).

The predictions by the LMS for this cluster are relatively accurate. Car driver use is still overestimated and walking underestimated, but the errors do not show irregularities. Cycling is overestimated, similar to cluster 2 where cycling was underestimated. Especially the LMS predictions for BTM and Train seem accurate.

- **Cluster 6 and 3: Older towns/ suburbs and suburbs of medium-sized cities**

  The zones in these two clusters are mostly zones around the medium-sized cities or suburbs of large urban areas and have a DU ranging from 3-5. Because of the large similarities between the clusters, they will be analysed together. This will also help with the argument to keep them separate clusters. (When using 6 clusters, these two clusters were merged, meaning that they are the two most similar clusters, according to the hierarchical clustering technique.)

  The relationship between the clusters 6 and 3 is similar to the relationship between the clusters 2 and 5 (the medium sized cities and the suburbs of large urban areas). The densities of these clusters are similar, though the population densities are slightly higher in cluster 3. The parking fee in these clusters is significantly lower than in the clusters 5, 2 and 1. Cluster 6 has a higher average parking fee than cluster 3. When looking at the violin plots, it shows that many zones in cluster 3 have no or a very low parking fee. The road density, however, is higher in cluster 3. Both clusters have few tram/ metro stops, although the number of stops in cluster 3 is slightly higher. This is because cluster 3 is more prominent in the zones around the suburbs of the large urban areas, where there are a lot of trams/ metros. The share of service land use has a similar distribution for the clusters 6 and 3, and is slightly lower than the share in cluster 5. The average number of bus stops is also slightly lower than in cluster 5, though the distributions in the violin plot shows that cluster 6 has more zones with a higher number of bus stops, giving a higher median. The distance to points of interest seems to increase from cluster 6 to 3, which follows the linear trend of increasing distances.

  The largest difference between the clusters is the share of houses built before 1945. In cluster 3 only a low share of houses is built before 1945, the lowest of all clusters, while cluster 6 has a relatively high share of older houses. This trend is confirmed by looking at the historical development of several locations where both clusters are present (e.g. Arnhem, Apeldoorn or Utrecht) using a website called *topotijdreis* (in English: 'topographic time travel', https://www.topotijdreis.nl/). This website makes is possible to look at maps of the Netherlands from different time periods and showed that the zones in cluster 6 often developed first (after the development of the city centres, i.e. cluster 2). After that, the city would expand to the zones in cluster 3.

  All these increasing or decreasing trends between the variables of cluster 3 and 6 are also seen between the clusters 2 or 5, except for the road density. The differences between clusters 2 and 5, however, are more pronounced in most cases.

  These parallels between the cluster pair 6 and 3 and the cluster pair 2 and 5, can also be seen when looking at the modal split. Again, the differences between clusters 6 and 3 are smaller, but they follow the same trends. Car driver use decreases with 1.5 %-points from cluster 3 to 6 when

looking at OViN, but the LMS predicts an increase in car use of 1.5 %-points. Train use increases from cluster 3 to 6 with 1 %-point, while BTM use slightly decreases. These trends are correctly modelled by the LMS. The relative differences between OViN and LMS are very large for BTM, but because the absolute values are small, the relative differences might not be the best way to compare OViN and LMS. Bike use in cluster 6 is also almost 3 %-point larger than in cluster 3, while the LMS predicts almost similar levels of bike use. The share of walking is very similar, though a bit lower in cluster 6. This trend is correctly modelled by the LMS, although the walking share itself is underestimated.

- **Cluster 4 and 0: Towns & small cities and rural areas**

  The zones in this cluster form the rest of the Netherlands. Cluster 4 forms the smaller cities and towns surrounding large and medium-sized cities, with mostly a DU of 3. Cluster 0 fills up the remaining of the country with mostly a DU of 1 or 2, and is classified as rural areas. These clusters are again analysed together, because the both follow a similar trend without large irregular observations.

  The density in both clusters is very low (though higher in cluster 4). The difference in density is larger when including the surrounding areas. This is true for both job and population density. This indicates that zones in cluster 4 are generally closer to more urban areas. The number of tram/ metro stops is very low in both clusters and the average number of bus stops and service land use drops almost linearly starting from cluster 3 up to cluster 0. The distance to points of interest increases almost linearly for the clusters 1 to 4, and make a significantly larger jump from cluster 4 to 0, indicating a larger travel distance to many necessities in the rural areas. The share of houses built before 1945 is larger in cluster 0 than in cluster 4

  When looking at the modal split, both OViN and the LMS do not show any unexpected results. Car use decreases from cluster 0 to 4, while the share of all other 4 modes increases. The errors in the LMS predictions follow the patterns identified in the earlier clusters (overestimation car driver use; underestimations walking; relatively large errors for BTM, though the absolute errors are small; and bike use that is modelled on a similar level on all clusters).

To conclude, this cluster set shows that it is possible to get clusters with very different travel behaviour and characteristics, even though the population density does not differ much. It seems that the modes popular in city centres (clusters 1 and 2), are also more popular in their respective suburbs. Cluster 1 has very high BTM use and a high share of walking, but an average level of cycling. Cluster 5, the suburbs of cluster 1, shows a decrease in all modes favoured by cluster 1 and an increase in car use. However, the favoured modes (BTM and walking) are still very high. Cluster 2 is unusual due to its high level of bike use, low BTM use, and large overestimation of car driver use by the LMS, especially when compared to the next cluster (cluster 5). Cluster 6, which could partly be classified as the 'older suburbs' of cluster 2, also shows the same patterns, though less extreme. The lower share of public transport and active modes in the clusters 5 and 6, are explained by an increase in car use. Cluster 3 can be seen as the 'newer suburbs' of the medium-sized cities, though it can also be spotted around zones of cluster 5. The final two clusters (4 and 0) show the same trends as seen when looking at the DU. This could indicate that in more rural areas, the Density is a good way to cluster zones. However, when the Density becomes higher, it is desirable to include more D-variables.

Analysis of unweighted cluster set

This section gives an analysis of the unweighted cluster set. This is done in less detail than the previous section, to avoid a lot of repeated information, because the clusters show a lot of similarities.

Table 4.5 shows the variables that were used to create this cluster set. This time, not all 6 D-variables are represented and the variables have no weights (i.e. they all have a weight of 1). Again, 7 clusters seemed to give the best results, contrary to the results from the different indices. The Davies-Bouldin score and the Calinski-Harabasz score favoured 2 and 3 clusters respectively, and the silhouette score 2. However, the silhouette score had a local maximum at 6 clusters, before making a significant drop from 6 to 7 clusters. It was still decided to use 7 clusters, because at 6 clusters a lot of detail was lost in the more rural areas. See appendix H for the values of the different indices and the dendrogram. Table 4.6 gives a small overview of the sizes and the name of each cluster. Figure 4.17 shows the different clusters on a map of the Netherlands, figure 4.18 shows how the different variables are distributed in

Table 4.5: Overview of different variables used to create the unweighted cluster set

D-variable	Variable
Density	Population density Population density, including surrounding zones Job density, including surrounding zones
Diversity	Share of service land use
Design	Road density
Distance to transit	Number of tram/metro stops
Demand management	Parking fare

Table 4.6: An overview of the 7 clusters with the names and the size of the clusters from the unweighted cluster set.

Cluster number	Cluster name	Number of zones	Share of zones
1	Centres of large urban areas	48	3.4 %
4	Inner suburbs of large urban areas	35	2.5 %
2	Centres of medium-sized cities	43	3.1 %
0	Outer suburbs of large urban areas	95	6.8 %
3	Suburbs medium-sized cities	348	24.8 %
6	Towns & small cities	244	17.4 %
5	Rural areas	593	42.4 %
Total		**1406**	**100 %**

each cluster, figure 4.19 shows the differences between these clusters and the DU and figure 4.20 shows the modal split for the clusters. The clusters are sorted based on car driver use according to OViN.

Next, an analysis of each cluster is given.

- **Cluster 1: Centres of large urban areas**

  Similar to the weighted cluster set, the first cluster consists of the city centres of Amsterdam, The Hague and Rotterdam and is fully made up of zones with a DU of 6.

  This cluster scores the highest on all three Density variables, has the highest average parking fee and the highest number of tram/ metro stops. The road density is high, but not as high as in cluster 4, the inner suburbs of large urban areas. The share of land used for services is significantly lower than in cluster 2, the centres of medium-sized cities. None of these variables seem out of place.

  The modal split for this cluster is similar to the corresponding cluster from the weighted cluster set: Low car use, a high share of public transport, and walking. It should be noted, however, that the share of walking is high, but not the highest from all clusters. The predictions from the LMS are also in line with earlier observations (fairly accurate public transport modelling, car is severely overestimated, and walking underestimated).

- **Cluster 4: Inner suburbs of large urban areas**

  The so-called inner suburbs are primarily the zones around the cities of Rotterdam and The Hague. Amsterdam is mostly surrounded by cluster 0, that is classified as the outer suburbs of large urban areas. The cluster is very small and consist of zones with a DU of 6.

  An important distinction between the zones of this cluster and the zones of cluster 1, is the job and population density that include the surrounding areas. These densities are considerably lower, while the population density of the zones themselves is on an almost similar level. This indicates that the zones of inner suburbs are closer to the borders of the city, than the city centre cluster. Both the parking fee and the share of land used for services is relatively low, compared to the three densest clusters. The road density is the highest of all clusters and the number of tram/

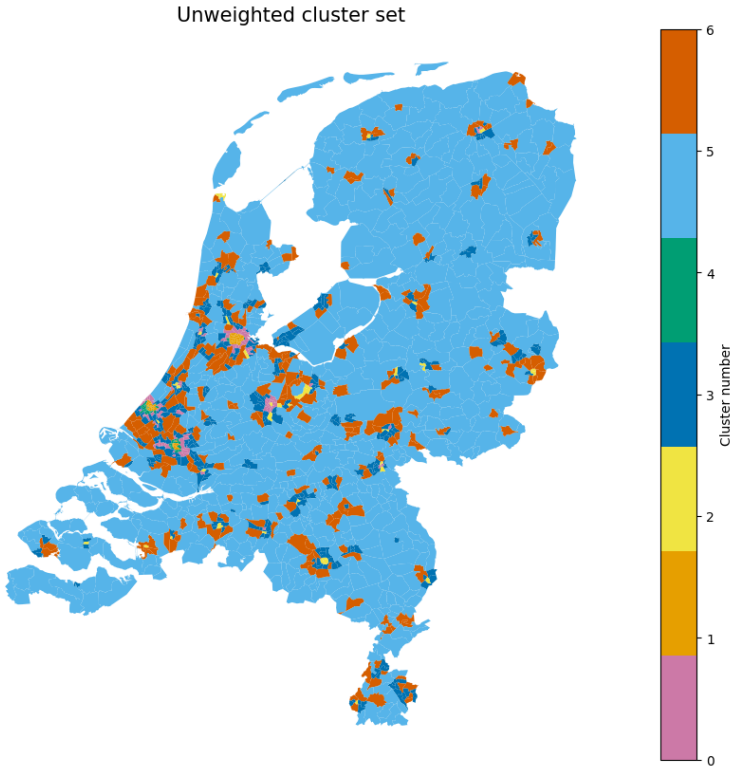

Figure 4.17: Map of the Netherlands displaying the unweighted cluster set. This map is created using hierarchical clustering. For the variables used to create this cluster set see table and 4.5. The shapes the zones are based on RWS WVL (2020).

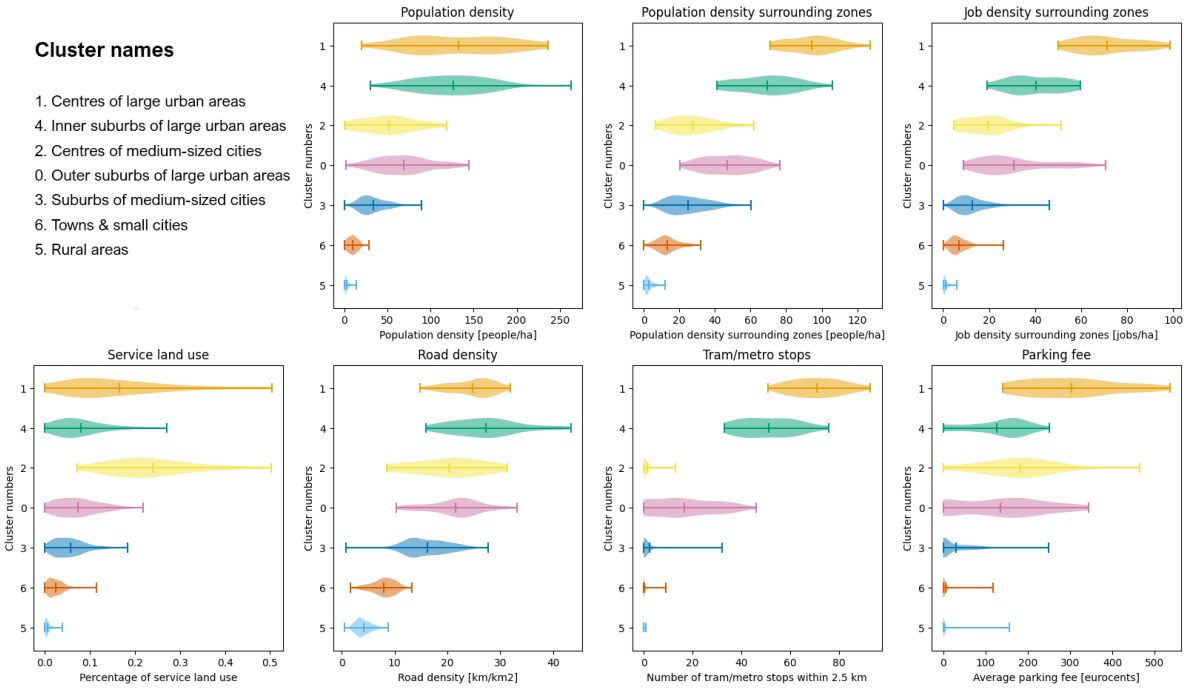

Figure 4.18: Distribution of the variables in each cluster for the unweighted cluster set in a violin plot. This figure is based on RWS WVL (2020) and the variables from table 4.5.

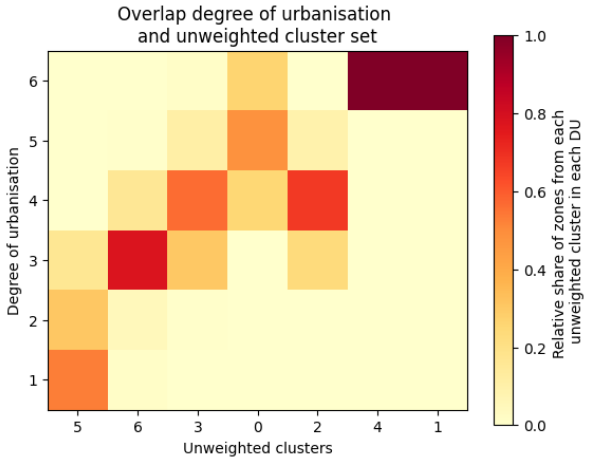

Figure 4.19: Comparison of the degree of urbanisation with the unweighted cluster set. Each square shows the share of zones of a certain cluster, belonging to a certain DU. For example, of the zones in cluster 2, around 70% has a DU of 4, around 20% a DU of 3 and around 10% a DU of 5. This figure is based on RWS WVL (2020) and the variables from table 4.5.

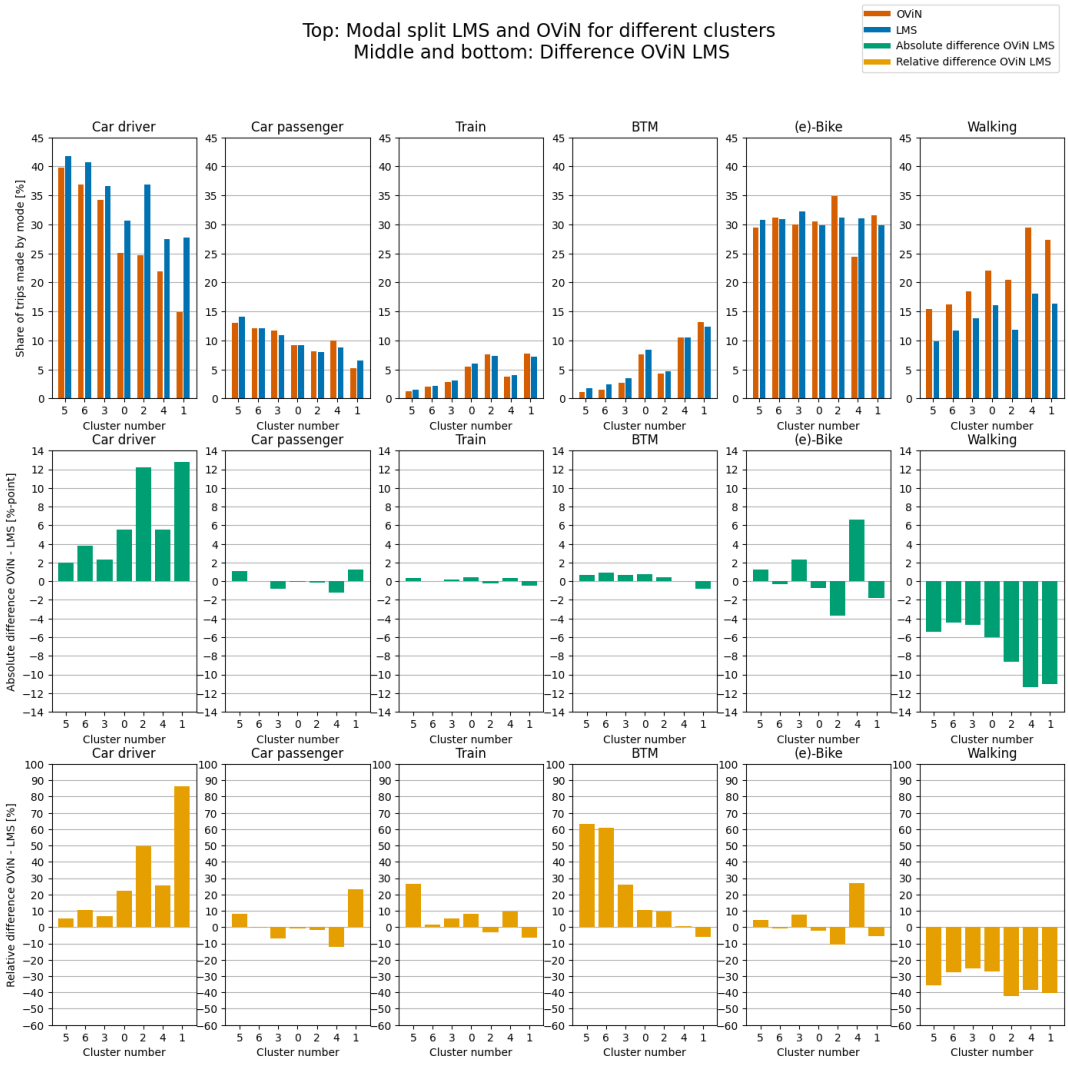

Figure 4.20: Top: Comparison between the modal split for different clusters of the unweighted cluster set for OViN (red) and LMS (blue). The percentages are based on the number of trips. Middle: Difference in %-points between OViN and LMS (OViN minus LMS). Bottom: Relative difference between OViN and LMS (difference = (LMS - OViN) / OViN). This graph is based on the combined OViN dataset from 2013-2017, the LMS OD-matrices (RWS WVL, 2018c) and clusters from the unweighted cluster set, which are based on the variables in table 4.5.

metro stops is also very high. This indicates that this cluster has a high quality public transport network. It presumably has a higher share of residential land use and is less tourist oriented than cluster 1. (Lower share of services and a lower parking fee, indicating that cars are more welcome in this part of the city.)

The modal split shows that the both the share of car drivers and the share of car passengers is significantly higher than in cluster 1. Car passenger use has not been mentioned often in the review so far, because it mostly shows a very predictable trend (increases when the share of car driver increases, though the differences between regions are fairly small. The LMS predictions are also good). However, in this cluster car passenger use shows a small peak, which has not been seen in the previous cluster set, nor the modal split based on the DU. Another interesting result is the low level of train use. This dip was not shown as clear in the weighted cluster set and is interesting, because a DU of 6 is associated with a high train use. Similarly to cluster 5 (the suburbs of large urban areas) in the weighted cluster set, the use of bike is significantly lower than in the rest of the country, while BTM use is very high. The share of walking is the highest of all clusters at almost 30%.

The LMS shows some interesting differences with the OViN results. First of all, car driver use is again overestimated, but also estimated to be on a similar, or even slightly lower, level than cluster 1. This could imply that the LMS predicts similar levels of car use for all zones with the same DU. This hypothesis is strengthened by the earlier analysis focusing on Amsterdam, which showed similar predictions of car use in the whole city. Cycling is largely underestimated, due to the extremely low bike use. The share of walking is overestimated.

- **Cluster 2: Centres of medium-sized cities**

  This cluster contains the centres of medium-sized cities, similarly to cluster 2 from the weighted cluster set. An important difference, however, is that this cluster contains half the zones compared to the weighted cluster set. This can be seen by comparing the somewhat larger cities (e.g. Groningen or Utrecht), where cluster 2 from the weighted cluster set identifies larger city centres. The zones of this cluster mostly have a DU of 4 or 5.

  All Density variables for this cluster are significantly lower than those of the clusters 1, 4 and 0 (the centre and suburbs of large urban areas). The parking fee, however, is high and the share of land used for services is the highest of all clusters. The number of tram/ metro stops is close to 0 and the road density is above average.

  The modal split is similar to the parallel cluster from the other cluster set: Fairly low car and BTM use. The highest share of train and bike and a share of walking that is slightly lower than in the denser areas. LMS again overestimates the share of car driver, and models it to be higher than car use in cluster 0, the outer suburbs (which has a similar car use as cluster 2 according to OViN). It models car use to be on a similar level as cluster 3, the suburbs of the medium-sized cities, which is made up of zones with a similar DU as cluster 2. Cluster 2 differs from cluster 3 through a significantly higher service land use and parking fee. The population and job density is also higher, though the distribution of the surrounding population density is fairly similar. This cluster pair is a good example of two clusters with a similar DU, but different characteristics of the spatial environment that show significantly different travel behaviour.

  BTM and train is modelled fairly accurate by the LMS, bike use is underestimated due to the high peak, and the share of walking is underestimated again.

- **Cluster 0: Outer suburbs of large urban areas**

  This cluster is mostly made up of zones around the four largest cities (Amsterdam, The Hague, Rotterdam and Utrecht) and a few zones around medium-sized cities. The zones in this cluster are mostly made up zones with a DU of 5 (and to a lesser extent 4 and 6).

  The Density variables in this cluster are higher than cluster 2, but lower than cluster 4 and 1, indicating that these suburbs are places farther from the dense city centres. The parking fee and road density is fairly high and the share of land used for services is on a similar level as cluster 4. The number of tram/ metro stops is significantly lower than those in clusters 1 and 4, but high in comparison with the rest of the clusters.

When looking at the modal split, the relationship between cluster 0 and 2 is to some extent comparable to the relationship of the clusters 5 and 2 in the weighted cluster set: The share of car drivers in both clusters is comparable, but the LMS predict lower car use in cluster 0 than in cluster 2. However, contrary to the modal split of cluster 5 from the weighted cluster set, the share of bike use is around the countries average. The bike share from cluster 4 from the unweighted cluster set is more comparable to cluster 5 from the weighted set. Interestingly, train use in this cluster is higher than in cluster 0, while BTM use is lower. A possible cause could be that the zones in this cluster are too far from the city centre to enjoy an equally high level of service for BTM, but are closer to train stations that can take them to the city centre. The share of walking is slightly higher than in cluster 2, and again underestimated by the LMS.

- **Cluster 3, 6 and 5: Suburbs medium-sized cities, Towns & small cities and rural areas**

  The final three clusters will be analysed together. This is done because these final three clusters do not show any irregular trends, but are in line with insights from the literature and the previous analyses. Cluster 3 consists primarily of zones with a DU of 4 (and to a lesser extent 3 and 5), cluster 6 consists of zones with a DU of 3 (and 4) and cluster 5 is made up mostly of the DUs 1 and 2. Cluster 3 is made up of neighbourhoods and small cities surrounding medium-sized cities and is classified as the suburbs of those medium-sized cities. Cluster 6 consists of zones around cluster 3, and includes smaller cities and towns farther from the larger cities. Cluster 5 makes up the rest of the country and is classified as rural.

  All variables follow an almost linear decreasing trend, from cluster 0 to 3 to 5 to 6, without notable exceptions. This is also seen in the modal split. Car use increases from cluster 3 to 5, Train, BTM and walking decreases and cycling stays approximately the same. The absolute predictions from the LMS are pretty accurate for all modes except car driver and walking, which follows the usual trend. It should be noted that the predictions for the share of car driver are relatively accurate, especially compared with the more dense areas.

To conclude, this cluster set managed to create different clusters that have the same DU (clusters 1 and 4 and clusters 2 and 3), while still having very different travel behaviour. Especially for the share of car drivers, the LMS seemed to give similar predictions to clusters with a similar DU, while according to OViN, there were very large differences.

Again, these clusters showed that when a cluster has a high share of train use, it doesn't mean that the BTM is also high, and the reverse. This is a trend that was not seen when looking only at the DU.

Comparison cluster sets and degree of urbanisation

Overall, when comparing both clusters sets, each has their own strengths and weaknesses. See figure 4.21 for a heatmap comparing the two cluster sets.

The unweighted cluster set is able to capture the differences in travel behaviour in high density areas with better detail. The inner and outer suburbs of the large urban areas show interesting differences in travel behaviour (bike and train use). In the weighted cluster set, these two clusters are partly merged. Although a large part of the zones in the inner suburbs, is part of the large urban city centre cluster from the weighted cluster set, and part of the clusters from the secondary suburbs, belongs to the medium-sized city centres from the weighted cluster set.

The unweighted cluster set, however, has several very small clusters, e.g. 3 of the clusters each contain less then 5% of the total number of zones and 4 are below 10%. The weighted cluster set also has 4 clusters that contain less than 10% of the zones, but only 1 cluster is below 5%. Larger clusters can make the overall results a bit more reliable and prevent overfitting.

The weighted cluster set seems to be better at differentiating between the different less urban clusters and more detail can be found outside the Randstad. This is an important quality, because the LMS is used to model transport for the whole country and not only the large urban areas.

Figure 4.22 shows the distribution of the spatial environment variables for the DU, based on the variables of the weighted cluster set. All variables in the unweighted cluster set are also part of the weighted cluster set. When comparing these new clusters with the DU, there are some interesting differences (see figures 4.4, 4.16 and 4.20 for the modal splits):

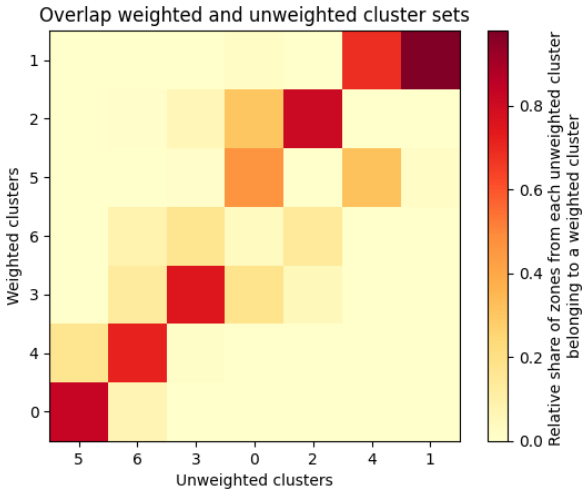

Figure 4.21: Comparison of the overlap between the unweighted and weighted cluster set. Each square shows the share of zones of the unweighted cluster set, belonging to a cluster from the weighted cluster set. For example, from the zones belonging to cluster 4 from the unweighted cluster set, around 70% belong to cluster 1 from the weighted cluster set and around 30% belong to cluster 5 from the weighted cluster set.

- When looking at the distributions of D-variables for the different DUs, several similarities can be seen with the distributions of the cluster sets (figure 4.14 and 4.18). For example, the average distance to points of interest decreases with a lower DU and with clusters that show lower car dependency. The correlation matrix (figure 4.11) already showed that this variable is highly (negatively) correlated with the Density variables.

  Other variables like the ratio of houses built before 1945 show very little correlation with the DU. The variables service land use and parking fee also show distinct differences when comparing the cluster sets with the DU. These variables seemed to play an important part in the creation of the clusters. This does not mean that the other D-variables are unimportant. None of the variables are 100% correlated with the DU, which means they all played a part in capturing the differences in the spatial environment and creating the clusters.

  Still, the DUs are able to show different distributions for most D-variables, which further strengthens the choice to use the DU as a proxy variable.

- It seems that the LMS is better in estimating higher shares of mode use, compared to lower shares. This becomes the most obvious when looking at BTM. The absolute differences between OViN and LMS are low, but the relative differences for the lower DUs or clusters with low BTM use are very high. Car driving shows a similar patterns. While the relative and absolute differences between OViN and LMS are very large with the higher DUs and clusters with a low car share, the differences become less when the share of car travel increases.

  A possible reason for this could be the limited amount of data that was available to train the LMS. This increases the chance of overfitting, which must be avoided. It also means that with those lower shares of a certain mode, the OViN results are less reliable because they are based on a limited number of data points.

- In general, the LMS seems to be fairly accurate in modelling the shares of car passenger, train and BTM. The absolute differences between OViN and the LMS are small for both cluster sets and for the DUs. This pattern was also seen during the clustering process, where many different combinations of variables were used to create the clusters. For all sets of zones that were clustered, the average predictions for the aforementioned modes were relatively accurate. Especially BTM and train showed interesting trends in specific clusters, and those trends were captured correctly by the LMS.

  The share of walking should also be mentioned. Even though the LMS always seems to underestimate the number of trips by foot, the different trends between the clusters were captured very

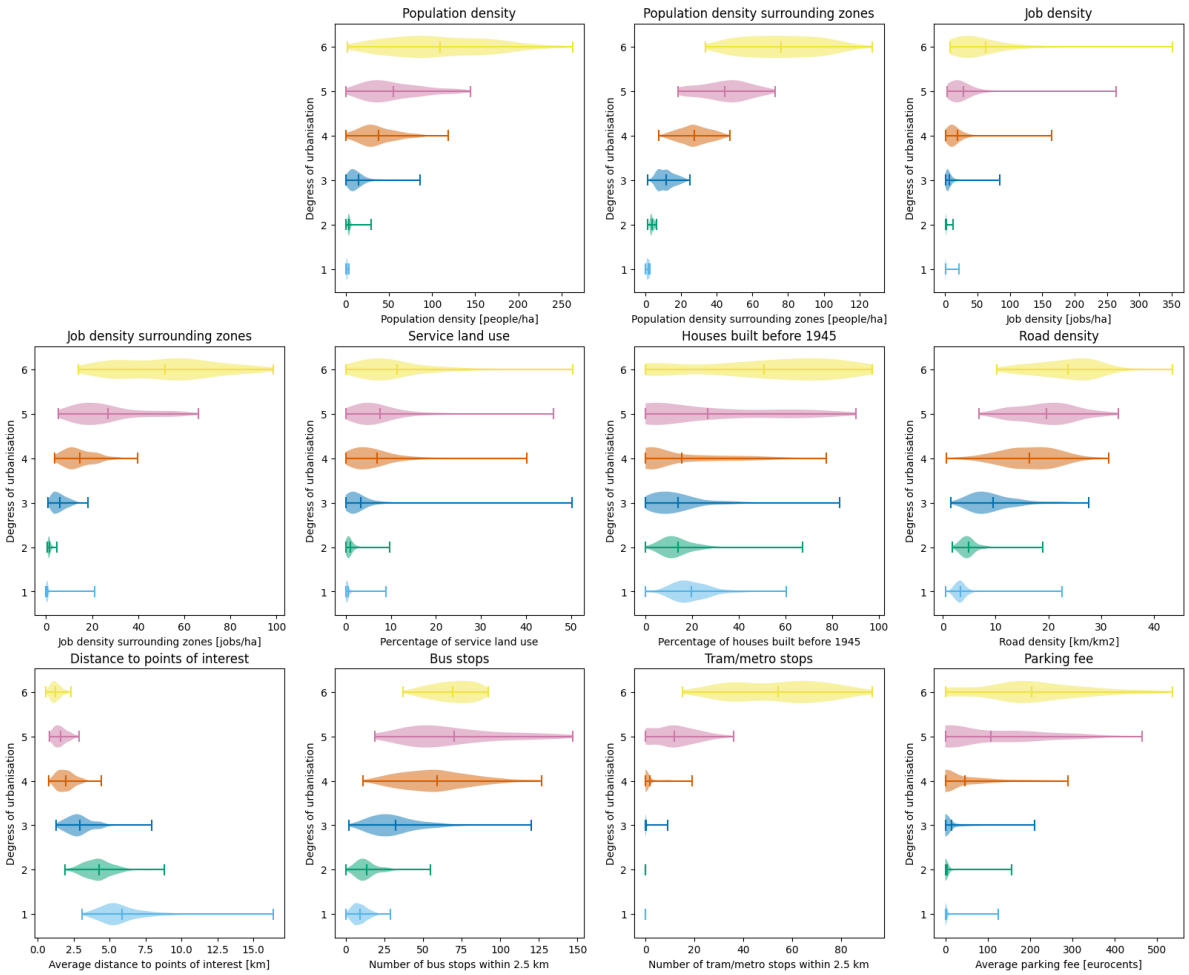

Figure 4.22: Distribution of the variables in each cluster for the degree of urbanisation in a violin plot. This variables are the same variables used in the weighted cluster set. This figure is based on RWS WVL (2020) and the variables from table 4.3.

well. When looking at both cluster sets, the same highs and lows can be seen. This implies that the LMS is able to capture the regional differences when modelling walking trips, but makes a mistake earlier when determining the overall frequency.

The different patterns over different regions by car drivers and bikes are on average captured the worst by the LMS. Especially the ‘suburbs’ and medium-sized city centres show interesting travel patterns that the LMS is currently not able to capture.

- Both the absolute and relative differences between OViN and LMS are smaller on average when looking at the DU, especially when looking at car drivers. This is a logical observation, because the LMS uses DUs in many different variables in their nested logit model.

  However, this also highlights one of the weaknesses in the LMS. Because the DU is an integral part of the LMS, regions with similar population densities, but otherwise very different characteristics, are joined together and are expected to have similar travel behaviour. This is very clear in the unweighted cluster set with cluster pairs 1 and 4; and 2 and 3. Both pairs have zones with similar DUs, and are modelled by the LMS to have similar shares of car driver use. According to OViN, the shares car driver use differ 7 and 10 %-points respectively.

To conclude, by capturing the differences between regions in different ways than only the DU, interesting travel patterns can be discovered that the DU alone is not able to capture. This could be a valuable addition to the LMS. However, before making recommendations, it is also important to control for the demography. This will be done in the next section.

4.4. Propensity score matching

In this section, propensity score matching (PSM) is done. The populations of the different clusters that were made during the clustering process have different demographic characteristics. Part of the differences in modal split between these clusters can presumably be explained by the differences in demographics. To make assumptions about the real effect of the spatial environment, clusters with similar demographic characteristics are compared. With PSM the demographics of each cluster pair are matched and the modal split of these matched clusters are compared. This will give better insights to what extent differences between the clusters are caused by the different demographics characteristics and to what extent the spatial environment affects travel behaviour. For more details about this method, see the methodology (section 2.5.3).

By comparing the results with the insights from the LMS documentation, some assumptions can be made about to what extent the differences between OViN and the LMS are related to the spatial environment and what parts have other causes. Besides that, advice can be given about which part of the LMS might require extra attention (e.g. if it appears that differences in modal split are mainly caused by differences in demographic characteristics, it means that modelling the accurate population of each zone should be a priority).

This section first describes the process of doing PSM and some of the choices that had to be made. After that, the results will be analysed. The PSM will be done for both cluster sets and for the DU, i.e. it will be done 3 times.

4.4.1. The process of propensity score matching

This subsection describes all the steps that were taken for PSM and the decisions that had to be made.

Comparison of the demographics

When looking at demographic characteristics of the clusters before doing PSM, there are large differences. See appendix J for tables showing the averages of each demographic characteristics for all clusters and DUs.

Choice of observations

For this thesis, there are few different ways to do PSM. As explained in section 2.5.3, observations with similar demographics need to be matched. First, the ‘observations’ must be defined. The different possibilities and decisions that were made are explained below.

- **LMS zones**

  All data (both the (D-)variables and the modal split data) is aggregated to zone level. This means that it is possible to treat each zone as an observation and match the zones of the different clusters with each other.

  This way, it is possible to compare the modal split for OViN and LMS with each other before and after matching. This is a large advantage of this choice of observation. However, with only 1406 different LMS zones, 7 clusters (or 6 DUs), and relatively large differences in demographics between clusters, it will be difficult to create good matches.

  This method was tested using 3 simple clusters based on 4 variables and 4 demographic characteristics, but it did not succeed. After testing, only a fraction of the zones were left and the standard mean difference (SMD) was often higher than 10% (see equation 2.5). Considering that the final PSM will include more clusters and more demographic characteristics, it was decided that using the LMS zones as observations was not possible.

- **OViN data**

  The second way to do PSM is by using individual OViN trips as observations. The advantage of this method is that the large number of data points makes it easy to find matches. However, this way only the OViN data can be matched, because the LMS OD-matrices does not include any demographic data but only the number of trips.

Even though the PSM cannot be used on the LMS data, the results from the OViN data can still provide valuable insights for the LMS (e.g. if the PSM shows that the spatial environment barely effects travel behaviour and the differences found in the previous sections are due to differences in demographics, implementing D-variables might not be the best way to improve the LMS).

After deciding to use the OViN data, additional decisions need to be made.

The first possibility is to use the individual persons in OViN as observations and then use all the trips each person made to compare travel behaviour. So far, travel behaviour of a cluster is defined by looking at the trips departing from that cluster. The destination cluster of a person might not be the same cluster as the origin cluster. By matching individual persons between clusters, this raises a problem. Do trips count that were started outside the home cluster or does a person count as a 'new' person when departing from a different cluster than their home cluster. To avoid this problem, each individual OViN trip counts as an observation and the trips are matched based on the demographics belonging to that trip. (In theory this means that when a person departs from two different clusters, it might be matched with itself and if a person makes multiple trips a day, it can also be matched multiple times.)

The final decision has to do with the weight factors. Each OViN trip has been assigned a weight factor. These factors range from 5.13 to 482.23 and indicate how many times a trip must be counted when making the travel behaviour representative for the whole country/ cluster. Due to the large range of factor values, it has been decided to include the weight factors when doing PSM. For this, the different factors are first rounded to an integer, because only whole trips can be matched. By using the weight factors, the data set is artificially enlarged. This is important when performing a t-test to test if there are significant differences between clusters before and after matching. (This is explained later in this section.)

Before doing PSM on the real data set, some test were done to see if the method works and to help with the decisions described above (e.g. using the LMS zones or the OViN trips as observations). See appendix K for these test results.

Estimate propensity scores

The first part of the PSM is to calculate the propensity scores (PS). This is done by fitting a logistic regression model (Pot et al., 2023). To do this, the demographic data needs to be processed. See section 4.1.2 for the list of demographic characteristics that were used.

- Some of the demographic characteristics are categorical (e.g. household type). These characteristics need to be converted to dummy variables (i.e. a separate variable for each household type with a value of 0 or 1).

  When dealing with dummy variables, it is important to not include all variables from the same category, but to have one reference variable (e.g. there are 4 different household types distinguished, but only 3 will get a dummy variable) (Taboga, 2021).

- It is not possible to do logistic regression on a data set that containes missing values. All trips with missing values have to be removed, which is around 5 % of the data. It is assumed that the deleted trips did not significantly affect the results.

Estimating the PS was done for each cluster pair within both cluster sets and for each DU pair. The PS is the probability of an observation belonging to the 'treatment' group. (One cluster within a cluster pair is called the control group, and the other is the treatment group.) There should be enough overlap between the scores of the two clusters to create the matches. Appendix J shows the distributions of the PS for each cluster pair in each cluster set. These graphs show that there is enough overlap between the demographics of the clusters to do PSM.

Matching the observations

The second part of PSM is to match the observations from the different clusters. The previous section showed that there is enough overlap between the clusters demographics to do PSM. Observations of which the difference (caliper) between the PS was less than 0.01 can be matched.

From the smaller cluster of the cluster pair, generally between 80-100% of the observation made it to the final cluster, with a minimum of 63% (Matching of DUs 1 and 6). From the larger cluster, the percentage of observations kept varies a lot. This is because the original sizes of the clusters were very different. At lowest, less than 6% of the observations of a cluster is kept. (Cluster 5 & 4 from the unweighted cluster set). This seems very little, but it is still around 8000 unweighted observations (and over 800 thousand when including weights).

The next step is to calculate the SMD to see if the matching is successful. Figure 4.23 shows the SMD values before and after matching for each cluster pair and each demographic characteristic for the weighted cluster set. All of the SMD values after matching are below the threshold of 10%, even when the SMD values before matching are very high (e.g. cluster 0 and 1). The figures for the unweighted cluster set and the DUs can be found in appendix J.

4.4.2. Calculate the average treatment effect

After the cluster pairs have been matched, the difference before and after matching can be calculated and compared. The average treatment effect (ATE) is the difference in mean travel behaviour between two clusters after PSM. The ATE is considered to be the impact of the spatial environment. The observed effects (OBE) are the differences in mean travel behaviour before PSM. The difference between ATE and OBE represents the impact of the demographics on travel behaviour. The impact of the spatial environment can be expressed as the ratio ATE to OBE.

Appendix J shows the full results from the PSM for both cluster sets and the DUs. The next subsection analyses these results.

Calculate p-values

It is important to check if the differences in travel behaviour within a cluster pair are significant. This can be done by calculating the p-value using a t-test.

However, one of the inputs when calculating the p-value is the size of the dataset. By using weights for the OViN trips, the dataset is artificially enlarged. When calculating the p-values for these enlarged datasets, the p-value was in almost all cases lower than 0.05, and the differences were assumed, maybe falsely, to be significant.

The problem is mitigated by randomly drawing x trips from the enlarged dataset, where x is the size of the dataset without weights. From this sample the p-value is calculated. This process is done 1000 times and the number of times the p-value is larger than 0.05 is counted. If more than 5% of the tests fail, it is assumed that the differences in travel behaviour between the clusters is not significant. The results are presented in appendix J. The ATE and OBE values are removed if the differences are not significant. If the ATE or OBE value was missing, logically the ATE to OBE ratio is also removed. It is important to note that it does not mean that the two clusters are similar if p > 0.05, it only means that it cannot be said if the observed differences between the clusters are due to chance or if the clusters are really different (James et al., 2023, p. 77).

4.4.3. Analysis of matched clusters

This subsection analyses the results from the PSM. As said in the previous section, the full results can be found in appendix J.

General results PSM

This subsection will give a general overview of the results.

The ATE-OBE ratio (or ratio, for short) is a useful indicator to get an idea of the true effects of the spatial environment. It should be used with caution, because when the differences between clusters are small, the ratio can get, perhaps falsely, very large. By filtering clusters without significant differences, inflated ratios are mostly filtered. For these cluster pairs, it can still be interesting to look at the individual ATE and OBE values to see if the clusters are significantly different before or after matching. This subsection will often use the average ratio for a certain mode. This average value needs to be used with care, because cluster pairs were the differences were not significant before or after matching, are not included in this average ratio. If they were still included the averages would presumably be smaller (e.g. the OBE value had p < 0.05, while the ATE value had p > 0.05. In reality, the ratio would

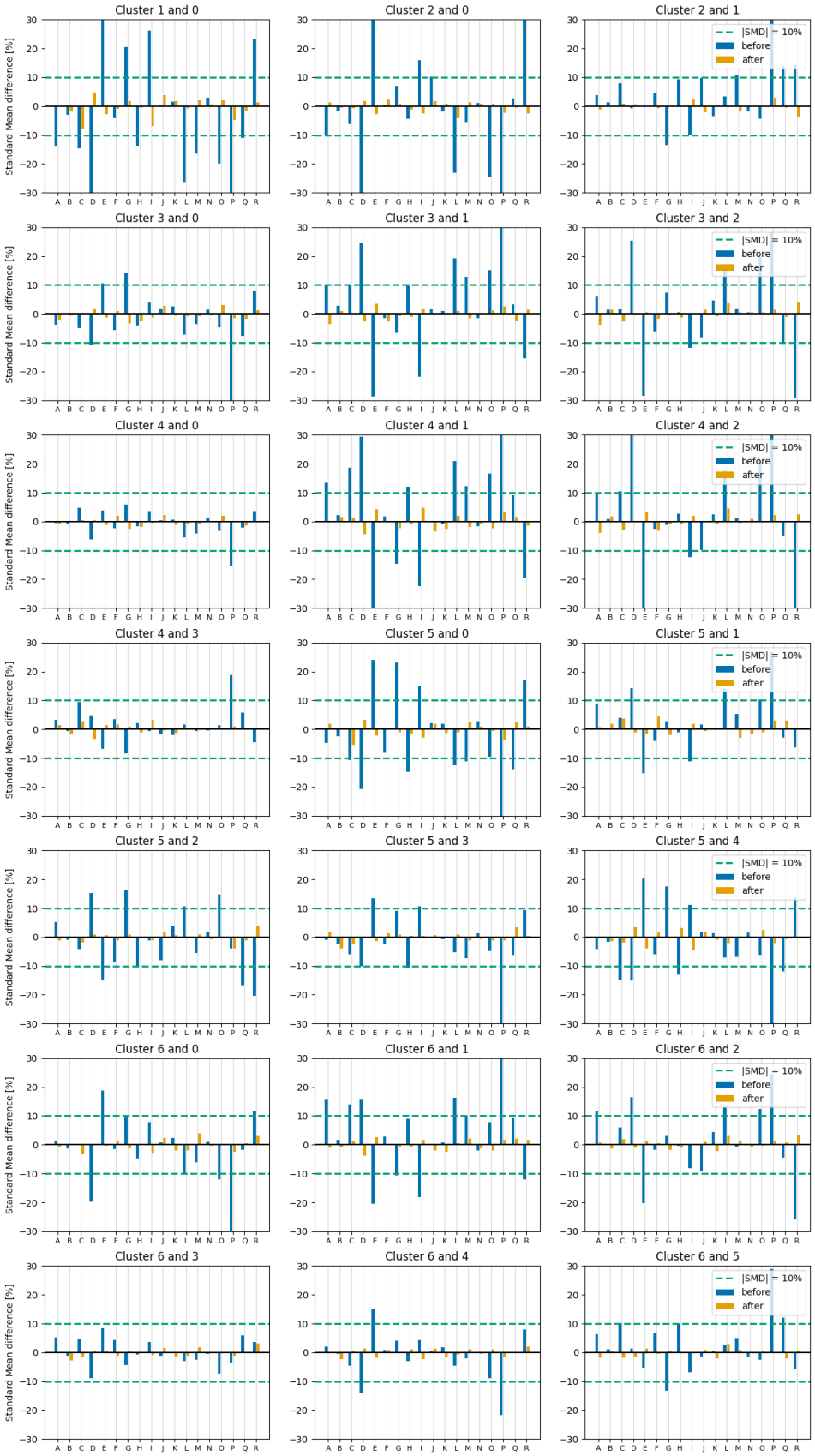

Figure 4.23: SMD values before and after performing the PSM for each cluster pair and each demographic characteristic for the weighted cluster set. The following demographic characteristics can be seen on the x-axis: A: Age; B: Gender; C: Income; D: Household size; E: 1 person household; F: 2+ person household; G: 1 parent household; H: Part time worker; I: Full time worker; J: Student; K: Primary education or less; L: lbo, vmbo; M: mbo, havo, vwo; N: other education; O: Younger than 15; P: Number of household cars; Q: Driver's licence; R: Student OV. The reference characteristics (e.g. 2 parent household for the household categories) are not included. All demographic characteristics are based on the combined OViN dataset from 2013-2017

Table 4.7: Average, minimum and maximum values of the ATE to OBE ratio for both cluster pairs and the degree of urbanisation.

	Weighted cluster set			Unweighted cluster set			Degree of urbanisation		
	Average	**Min**	**Max**	**Average**	**Min**	**Max**	**Average**	**Min**	**Max**
Car driver	0.65	0.34	1.00	0.62	0.33	1.02	0.51	0.39	0.69
Car passenger	0.60	0.45	0.85	0.61	0.34	1.06	0.51	0.31	0.64
Train	0.56	0.35	0.78	0.55	0.33	0.75	0.49	0.32	0.69
BTM	0.81	0.5	1.23	0.83	0.47	1.33	0.78	0.54	0.90
Bike	1.03	0.58	1.51	1.05	-0.99	2.34	1.07	-2.03	3.34
Walking	0.83	0.52	1.13	0.81	0.54	1.27	0.83	0.76	0.96

presumably be very small in this case and lower the average, but there is not enough evidence to give a reliable value for that ratio). See table 4.7 for the average, minimum and maximum ratios.

An important disclaimer is that all statements that are made about the effect of the spatial environment, are under the assumption that the matched populations are similar enough, and that all relevant demographic characteristics were included in PSM.

First, the results for car drivers are analysed. The car driver ratio is significant for almost all cluster pairs for both cluster sets and the DUs. For the weighted cluster set, the ratios differ between 0.34 and 1.00, with an average of 0.65. This means that the observed differences in the share of car driver trips between clusters can be explained for 34 up to 100 % by the spatial environment. For the unweighted cluster set the ratios differ between 0.33 and 1.02, with an average of 0.62 and for the DUs the ratios differ between 0.39 and 0.69 with an average 0f 0.51. A ratio higher than 1 (as seen in cluster pair 2 & 3 for the unweighted cluster set for car drivers), can be an indication that the effect of the spatial environment is underestimated, which means that there could be negative residential self-selection (J. Liu et al., 2024). In other words, when the exact same population would live in cluster 2 and 3, the differences in car travel would presumably be larger than is currently observed.

It is important to not skip the cluster pairs with no significant ratio. In some cases there are no significant differences between the cluster pairs before and after matching (e.g. clusters 0 & 4 for the unweighted cluster set); in other cases there are significant differences observed before matching, but they disappeared in similar populations (e.g. clusters 3 & 4 for the weighted cluster set); in the last case, there were no significant differences between two populations before matching, but they appeared after matching (e.g. clusters 2 & 0 of the unweighted cluster set). This is another case where the effects of the built environment were underestimated when demographic characteristics are not included.

When comparing the new cluster sets with the DU, it is interesting to see that the differences in car driver use between the DUs is only for about 50% due to the spatial environment, with limited variation between ratios. With the new cluster sets this ratio is larger on average, indicating that the new cluster sets are better in capturing the effect of the spatial environment. The new cluster sets do show a wider range of ratios, which could indicate that there is a lot of difference in the ‘quality’ of the clusters (i.e. how well a cluster is able to distinguish different type of areas). However, before more conclusions can be made, the other modes should be analysed. A cluster pair that shows little difference in one mode use, can still have large differences between other modes.

For car passengers, there are more cluster pairs without an ATE to OBE ratio. In most cases, no significant differences are observed after matching. However, in quite a few cases, there were already no differences before matching. The average ratios are slightly lower than the average ratios for car drivers for the cluster sets and the average ratio is the same for the DUs. Based on this, it seems that the choice to travel by car as a passenger is less affected by the spatial environment, than choosing to drive the car. A possible explanation for this could be the motives of the trips. Car passengers might have a higher chance of having a leisure motive or a higher chance of being a child (e.g. it can be perceived as more comfortable to travel by car when transporting several kids, regardless of the location), while car drivers might have a work related motive more often, which makes it easier to switch modes depending on the location. This has not been researched in this thesis.

Similar to the car drivers, the car passenger ratios for the DU pairs are lower on average with less

variation. However, due to large number of missing ratios, it is more difficult to make conclusions about the quality of the clusters.

For train users, there is a ATE to OBE ratio for most of the cluster pairs. Interestingly, the average ratios are the lowest of all modes. This is true for both cluster sets and the DU, though the average ratio for the DUs is lower than for the cluster sets. In the case of missing values, this was always due to no significant differences after matching (and sometimes also before matching), but not due to negative residential self-selection. This result can seem somewhat counter intuitive. It would be logical to assume that train travel is affected by the spatial environment for a large part, due to differences in quality and availability of the train network throughout the clusters. These results indicate, however, that on average almost half of the differences in train travel between the clusters can be explained by personal preferences.

The first noticeable observation when looking at BTM is that the effect of the spatial environment is very large, around 80 % for all cluster sets/ DUs. For the cluster sets there are 0 missing values before matching (all differences were significant), and 3 or 4 missing values after matching. For the weighted cluster set, all missing ATE values had already very low OBE values and a low number of BTM trips overall. For the unweighted cluster set, there are larger differences that disappear after matching (e.g. cluster 4 & 0, the inner and outer suburbs of large urban areas show a difference of almost 3 %-points before matching). When looking at the DU-pairs, no significant differences in BTM use could be found when looking at the lowest DUs. These DUs have a low share of BTM use.

There are several cluster pairs that show negative residential self-selection. Cluster pairs with high ratios, have at least one cluster with very high share of BTM most of times. Those are clusters that have a high quality BTM network. This indicates that the effect of the spatial environment is larger when there is high quality BTM. When neither cluster has a very good BTM network, even if the network in one of the clusters is significantly better, the effect of the spatial environment is smaller and people's preferences start playing a larger roll. For example, cluster 2 & 0 from the weighted cluster set (medium sized city centres and rural) both have low uses in BTM. The density of the bus network in cluster 2 is the highest of all clusters and in cluster 0 the lowest. However, there are almost no tram/ metro stops in both clusters. The differences in BTM use between cluster 2 & 0 is for only 50 % determined by the spatial environment.

Perhaps when there is only a bad or acceptable quality of BTM, people are more inclined to use the other modes. These modes might be more convenient than using the bsu, even when there is a clear improvement in the bus network. Most of the people left on the bus would be the people who prefer riding the bus or have no other choice. In dense urban areas with a high quality BTM network, using the BTM can be more convenient than other modes, so people who would normally take other modes, are now more inclined to use BTM (i.e. BTM start being a real competition to other modes when the quality reaches a certain level). This is a possible explanation, but has not been further research in this thesis.

The effect of the spatial environment on bike use, might be the most interesting of all modes. There are a lot of missing values, but the existing ratios have an average larger than 1 for all cluster sets/ DU, indicating negative residential self-selection. The OBE-values contain a lot of missing values, indicating that the differences were not significant. The ATE values, however, have a less missing values, especially when looking at the weighted cluster set and the DUs. In other words, where there were no significant differences in bike use before matching, they often appeared after matching. This strengthens the theory of negative residential self-selection. There are even several cluster pairs where the differences in bike travel more than doubles (e.g. cluster 1 & 2/ city centres large urban areas & medium sized city centres from the unweighted cluster set) or the effect reverses (e.g. DU 2 & 4, where there was more cycling in zones with a DU of 4 before matching, and more cycling in zones with a DU of 2 after matching).

These results indicate that when modelling bike use, it is extra important to include spatial variables to correctly capture the different trends in bike use throughout the country. When using primarily personal characteristics, places that tend to have a lot of bike users might be overlooked (e.g. the LMS shows largely underestimates bike use in medium-sized cities).

For walking, there are few missing values. For a few cluster pairs there were no OBE values and no ATE values or only no ATE values (i.e. it did not happen that differences only became visible after matching). The ratios for all cluster sets/ DU averaged above 0.8. This means that differences in walking shares between clusters were largely caused by differences in the spatial environment. There are a few cases with ratios higher than 1 and no ratios lower than 0.5. Interestingly, the ratios for the DU-pairs differ from 0.75 to 0.96. This could indicate that the DU is good way to separate zones when looking at walking behaviour. The cluster sets perform similar on average, but do have several cluster pairs with ratios close to 0.5. The hypothesis that the DU can be a good way to model walking behaviour in different zones, is strengthened when looking at the results from the cluster analysis. One of the conclusions was that the LMS seems to be good at predicting walking behaviour trends in different zones, regardless of which clusters were made. A possible explanation for these could be that the currently implemented DU works very well.

Finally, when looking at the ATE and OBE results of the DU, there are no significant ATE or OBE values when comparing the DUs 1 and 2. This means that even before matching the populations, there is no significant difference in the modal split for these regions. This is in line with the results from the cluster analysis, where both cluster sets put most zones with a DU of 1 or 2 in the same 'rural' cluster. On one hand, the DU might not differentiate enough between different type of urban regions, while on the other hand it might not be needed to differentiate too much between the rural areas.

For both cluster sets, there are no cluster pairs that showed no significant differences for all modes. This indicates that all clusters are relevant and are able to identify regions with different travel behaviour. There are, however, some clusters with only a few differences after matching. For example, the inner and outer suburbs from the unweighted cluster set only showed significant difference in the shares of cycling and walking. People from the outer suburbs seem to substitute walking for cycling, while other mode uses are similar.

Conclusions effect spatial environment

This subsection gives some conclusions based on the insights obtained from the ATE and OBE values.

- First of all, the results indicate that the spatial environment plays a large roll in the differences in modal split between different clusters. In most of the cases the spatial environment is responsible for more than 50% of the differences and in some case the effect of the spatial environment is underestimated, when the demography is not accounted for. This means that it is important to include variables with spatial characteristics in transport models. By using only (or primarily) demographic characteristics, it might not be possible to accurately model travel behaviour.

- Even though some cluster pairs showed no significant differences for certain modes before and/or after matching, no cluster pairs showed no significant differences at all after matching. In other words, each cluster is relevant. The only exception is the DUs 1 and 2, that are mostly combined in the cluster sets.

- The average ATE to OBE ratios for car drivers and passengers are higher for the cluster sets than for the DU. Where for the DU the spatial environment is able to explain around 50% of the differences in car use, the spatial environment in the cluster sets are able to explain more than 60% of the differences. For train and BTM, the cluster sets also have higher ratios on average than the DU.

  This could indicate that the new cluster sets are indeed better at differentiating regions with different travel behaviour than the DUs. In other words, it could be beneficial to include more D-variables in a transport model, than only Density variables.

  The current LMS predictions show a lot of errors when determining the share of car drivers, especially for clusters close to city centres. By using different clusters for which the differences in car driver use can be explained for a larger part due to the spatial environment, it might be possible to improve the predictions for car use.

- These results indicate that differences in BTM, cycling and walking between different clusters are caused mostly by the spatial environment (>80% on average). For these modes it might be extra

beneficial to add additional D-variables to a transport model to account for these differences in the spatial environment.

Differences in car and train travel are for a larger part due to personal preferences. For these modes it can be more important to make sure that the population distribution for each zone is realistic to account for the differences in modal split caused by different demographic characteristics.

Of all modes, the results for bike use are the most interesting, because of the negative residential self-selection. The exploratory and cluster analysis already showed that there can be a lot of difference in bike use regionally, while bike use is fairly constant on a national level. Only Density variables are not enough to explain those local differences in bike use. Bike use might benefit the most from adding additional D-variables to a transport model.

5

Discussion

This chapter discusses the results obtained in this thesis. First, there is reflected on the approach and the limitations of this study are discussed. After that, the results from this thesis are compared with the literature found during the literature review. Finally, the generalisability of the results is discussed.

5.1. Reflection on approach/ limitations

The limitations will be discussed in several steps, by reflecting on the approach of this thesis. First, the limitations with regard to the scope will be discussed, then the limitations with regard to the handling, processing and analysing of the data and finally the limitations of the results.

5.1.1. Limitations of the scope

This thesis focused on differences in travel behaviour between different regions. However, this thesis only analysed the differences in modal split, looking at all trips made throughout the day from a certain origin zone.

Initially it was the plan to include travel distance, travel time and part of the day in the analysis. However, this turned out to be more difficult than thought in advance. The OViN data included travel time and distance for each trip, which made it easy to include. For the LMS data, however, it was more complicated to obtain the travel time and distances, especially for public transport. For example, available data included travel time and distances separately for train and the access and egress. To capture the full distance and duration of a trip, it would require a lot of data processing. An estimate for travel time and distance was made for each mode using daily average travel times and distances according to the LMS. For train travel times, a general public transport travel time matrix from the LMS was used. The distances for train were estimated based on the relation between travel time and distance according to OViN. This was done using linear regression. The resulting graph can be seen in appendix G. The travel times and distances for OViN and LMS differed a lot. (The LMS showed higher travel distances on average, but lower travel times, indicating a higher travel speed. The latter could partly be explained by the fact that the LMS overestimated the number of trips by car. The average durations and distances were calculated together for all modes.)

It is unclear if there are incorrect assumptions in calculating the travel times and distances for the LMS; if distances and duration are measured in different ways in OViN and the LMS; or if the destinations predicted by the LMS are indeed very different from the results from OViN.

There is a relatively easy way to mitigate this problem and to analyse the differences in destinations between OViN and LMS. This can be done by using one similar way to calculate the distances and durations for both OViN and LMS. For example, by taking the euclidean distances between the origin and destination zones, insights can be obtained about the accuracy of the LMS in predicting destinations. Another possibility is to process the data from LMS in more detail, separately for different modes and part of days and compare travel times and distances for origin destination pairs for OViN and LMS to evaluate the accuracy. This is very time intensive.

To conclude, due to time restrictions and because determining the travel time and distances according to the LMS is more difficult than originally planned, the above solutions are not implemented in this thesis and can be done in future research.

The LMS includes the part of the day in their nested logit model. The original plan was to include this aspect of travel behaviour in the analysis. This was not done because of several reasons.

First of all, departure time is a subject barely touched in literature. Most of the literature focuses on the modal split and travel time and distances. Secondly, the LMS determines the part of day not for all modes, but only for car drivers and passengers, train and BTM. In other words, it is not possible to analyse the modal split during, for example, the morning peak. It would have been possible to analyse the modal split based on the other modes, but the active modes are responsible for about 50% of the trips. Analysing only the share of trips for the remaining modes, could lead to wrong conclusions. Appendix G shows some of the preliminary results when looking at the differences in modal split during different parts of the day for car and public transport.

There is also the travel frequency: the number of times a person makes a trip per day. The initial plan was to also include the travel frequency, but it turned out to be impossible with the available data. The LMS OD matrices only show the origin and the destinations of all trips, including the modes, but nothing was known about the people who made those trips. This made it impossible to calculate the frequency. In another part of the LMS, the travel frequency for each type of person is determined. When having that data, it would be possible to compare the frequencies of OViN and LMS.

A final aspect of travel behaviour is the travel motive: the reason a person makes a trip. This aspect was also not included in the analysis, due to limitations in the LMS data. In the available LMS data, the different travel patterns for each motive were only available for car drivers.

Besides looking at more aspects of travel behaviour, it could also be interesting to look at tours instead of trips. This is how the LMS models travel behaviour. The output of the LMS and the OViN data are both given in trips, so that was analysed for this thesis. By analysing tours instead of trips and including both characteristics from the origin and the destination, it might be possible to get a deeper understanding of how the spatial environment affects travel behaviour. On average, a person will use the same main mode for all trips in a tour, so it would be logical to assume that a person chooses a mode that is suitable for both locations. By looking at only at the origin or destination, valuable information affecting the mode choice might be left out.

If all the factors mentioned above would have been included in this thesis, a more in depth analysis would have been possible, giving more insights in the differences in travel behaviour between regions. Besides that, it would have been easier to find out the reasons for differences between OViN and LMS (e.g. travel patterns of one of the motives has large errors, while other motives are predicted well). However, this thesis still provides valuable insights about the effect of the spatial environment on the modal split. This research can serve as a starting point for more in depth analyses in the future.

5.1.2. Limitations of the literature review

This subsection covers the limitations of the literature review.

There is a lot of literature about the effect of the spatial environment on travel behaviour. Many papers were assessed, but due to the large number of papers and limited time, it was not possible to read everything. This makes it possible that relevant information was overlooked. However, as far as the author knows, a broad overview of the available literature has been obtained.

Based on the literature review, the D-variables were found as a way to quantify the spatial environment. For each D-variable, several variables were gathered that were assumed to capture that D-variable (e.g. for the D-variable Destination accessibility, the variables: ‘distance to several points of interest’ and ‘distance to city centre’ were gathered). However, there is no guarantee that these variables capture the full effects of the spatial environment.

5.1.3. Limitations of the data processing

Several limitations of this thesis can be identified with regards to the data and the data analysis. These limitations are discussed in this section.

The first step for processing the data was to match the PC4 zones with the LMS zones. Many times the borders of the LMS zones followed those of the PC4 zones, but this was not always the case. In a few cases, a PC4 zone consisted of several LMS zones, in which case a choice was made to which LMS zone the corresponding OViN trips were assigned. This had to be done because the OViN trips did not contain exact locations of the origin and destination, but only the PC4 location.

Because of these choices, it is possible that some of the trips did not belong to their assigned LMS zone, but actually to a neighbouring zone. These inaccuracies matter less when doing the cluster analysis, because often clusters included several neighbouring zones and all trips were aggregated when estimating the modal split for specific clusters. Of course, not all neighbouring zones belonged to the same cluster. Besides that, when zooming in and analysing differences in travel behaviour within a city (e.g. the analysis for Amsterdam), the exact differences in modal split between neighbouring zones could be less accurate.

To summarize, because of inaccuracies in matching the PC4 zones with the LMS zones, the modal split of each zone might also have some inaccuracies. These inaccuracies can become more significant when looking at only a small set of zones.

A possible way to partly deal with this problem in future research, would be to assign all trips to both zones, but divide the corresponding weight factors for each trip, based on area or population of both LMS zones. This way, it is avoided that some LMS zones get no trips at all. This method comes with the assumption that the trips are equally spread throughout the PC4 zone.

Similar problems were encountered when gathering the D-variables for the LMS zones. In this case it was possible to match each PC4 zone or each neighbourhood zones (a type of zone smaller than PC4) with multiple LMS zones. It was assumed that all PC4 or neighbourhood zones belonging to an LMS zone, fully belonged to that zone and that a LMS zone was fully made up of its assigned PC4 or neighbourhood zones. In other words, one PC4 zone could fully belong to two different LMS zones. This made it easier to calculate the averages for each LMS zone.

This assumption could have introduced additional inaccuracies in the variables for each zone. To give an extreme example, neighbourhood zone X belonged to both LMS zone A and B and no other neighbourhood zones were assigned to zone A and B. The land use for zone X was for 50 % residential and for 50% nature. In reality, it might be possible that zone A was a fully residential zone, while zone B was fully nature. However, with this method it was assumed that both zones were 50% residential and 50% nature.

Luckily, most of the LMS zones were accurately matched with the PC4 and neighbourhood zones, limiting the effect of wrongly assigned zones. Again, when zooming in on a small set of zones or individual zones, the D-variable characteristics might be less accurate. However, when looking at larger sets of zones, like the clusters, the D-variables are assumed to give an acceptable representation of the regional characteristics of that set of zones.

It was tried to gather D-variable data for 2018, because that is the base year for the LMS. However, not all data was available for that year. For example, most data on PC4 level (e.g. the corresponding shape files, data for distances to points of interest) came from a dataset from 2019. The data about land-use characteristics on neighbourhood level was from 2017. Data that was already available in the LMS data, was mostly based on data from 2018. Some data was even more recent (e.g. the ratio of bicycle and pedestrian roads was from 2022).

It was assumed that the data did not differ significantly from the base year of 2018. However, it is possible that there were some significant changes in the characteristics of some zones.

Finally, CBS data is censured when it contains less than 5 observations of a category that is associated with persons (e.g. houses with a certain built year) and all values are rounded to integers dividable by 5 (Van Leeuwen & Venema, 2023). This introduced additional inaccuracies, mostly in zones with a smaller population. It increased the chance that these zones were assigned to the wrong cluster. Luckily, this was only relevant for one variable that ended up in the final cluster sets (share of houses built before 1945). Besides that, the impact of these zones on the modal split of their corresponding

clusters were presumably small, because of the smaller population of the zones.

In other words, the chance that these zones belonged to the wrong cluster was higher, but their impact on the average travel behaviour was limited.

5.1.4. Limitations of the D-variables

Some of the D-variables were easy and straightforward to determine, without much room for inaccuracies (e.g. to calculate the population density, the total population of a zone was divided by the area. Both the population and the area were given by the LMS data). However, for some variables more assumptions had to be made. This section reflects on some of the variables and suggest ways for how they can be improved. Not all variables will be mentioned, but only those that did not make it to the final clusters, but were heavily represented in literature and variables that are assumed to have a lot of inaccuracies.

All variables discussed below are not included in the final cluster sets. All variables that made it to the final cluster sets, do not have a lot of known inaccuracies and gave logical results.

Entropy

The entropy index, a measure for the land use balance, is used a lot in literature (e.g. Ewing and Cervero, 2010; Kockelman, 1997; Limtanakool et al., 2006). Figure 4.12 shows that the correlation between entropy and the modal split lower than a lot of other variables. During the clustering, the entropy value was often the highest in smaller and lower density cities/ areas and considerably lower in the largest cities.

This is presumably related with the way the entropy was used in this thesis. Most of the studies of travel behaviour focused only on a large metropolitan area and used the ‘developed’ areas to calculate the land use (e.g. (Kockelman, 1997)). This study, however, also included land use types like agricultural and nature, which makes up a very large part of the land use in the Netherlands.

It might be possible that the entropy measure is less suitable when using it on a whole country with a lot of different type of regions, and more suitable when focusing on a large urban area, like done in literature.

Besides that, the literature also found that the entropy measure is negatively correlated with travel time and distance (Feng et al., 2013), these aspects of travel behaviour were not researched in this thesis.

Road width

One of the variables, corresponding to the D-variable Design, is the road width. This variable turned out to be very weakly correlated with the modal split. It is unclear if this is because the road width is indeed only weakly correlated with the modal split, or because of the inaccuracy of the variable. Due to a lot of missing data, not all road widths were given. It is possible that there was a bias in road widths that were missing (e.g. large roads like highways contained measurements, while small access roads in neighbourhoods did not). This was not further researched, because there was already another Design variable, that proved to be very effective (the road density). Besides that, the average road width was calculated, using all road sections in a zone according to the NWB shape file. However, the average width was not calculated using weights like the length of the road. This is something that could have been done better in hindsight.

The road width was a variable often found in papers based in the United States, which has a very different road network than the Netherlands. It might be good to develop new Design variables that are more suited for the Dutch road network. For a variable like Design of the road network, it might be extra important to look at the network for the whole trip, instead of looking only at the origin (or destination).

Proportion of bike and pedestrian roads

The final Design variable is the proportion of bike and pedestrian roads, compared to the length of all roads. The accuracy of this variable is also questionable. The dataset comes from 2022 and does only include separate bike paths, not painted bike lanes on shared roads, or bike paths that are shared with other slow modes (NWB, 2021).

The variable was not part of one of the final cluster sets. It might be valuable to use this variable again in the future, when the NWB has a more complete overview of all bike and pedestrian paths in

the Netherlands.

Another possibility for a future variable is to look at the connectivity of the cycle network to see if a zone has a connected cycle network instead of separate road sections.

Distance to city centre

As mentioned in section 4.1.2, the distance to the city centre is an important and useful variable according to the literature, but none of the sources gave an exact definition. This thesis attempted to create a method to determine the distance to the city centre for a whole country. Figure 4.12 showed that the distance to the city centre was correlated with the modal split (except for bike use), though the correlation was not extremely strong, compared to other variables. At the end, the average distance to points of interest seemed to perform better and was included in the weighted cluster set.

When looking at the maps from the unweighted and the weighted cluster sets (figures 4.17 and 4.13), it does seem like the distance to the city centre would have been a useful characteristic. Often, there are a few zones that are part of a 'city centre' cluster and several zones surrounding the city centre that also form a cluster.

When fine tuning the method for determining the city centre, it might be very useful. The current method stated that there could only be one city centre per municipality. However, looking at the current cluster sets, this might be too much. It might be better to have only one city centre per larger region, or a minimum distance between two city centres.

This variable might be a very valuable tool after improvements.

Distance to BTM and number of BTM lines

In section 4.3.1 it is mentioned that not all BTM related variables worked as intended. The 'distance to BTM' variable gave too little information about the quality of the network, putting variables within a certain distance of a tram/ metro stop automatically in another cluster. The number of BTM lines turned out to be a bad variable, giving no or little information about the quality of the BTM network. Both variables were replaced by a variable that counted the number of BTM stops within a certain radius.

The distance to BTM could maybe be improved by finding a better way to deal with zones that are very far from a tram/ metro stop, instead of giving them an arbitrary distance of 5 kilometres. However, it would still be important to combine this variable with another variable that says something about the quality of the network.

The number of BTM lines could maybe be improved by more data processing. The dataset used had the number of different BTM lines per platform, so large BTM stations with several platforms still scored low. Both directions of the same BTM stop were also counted separately.

This variable might be more useful when counting the number of different destinations that could be reached, by stops within walking distance of a certain zone. That way something can be said about how well the BTM network is connected, avoiding the problem of different platforms.

Besides that, obtaining data about BTM frequencies in a zone, could also be useful.

Parking places

For Demand management, the number of parking places in a zone were determined. However, the available dataset only contained parking places adjacent to roads. Parking garages or large parking lots were not taken into account. This resulted in a ineffective variable and the correlation with the modal split was low (figure 4.12).

This variable might be improved by finding a dataset that includes all parking places in a zone and divide that by the total population or the number of cars. This will show better how car friendly an area is. Besides the number of total parking places, there is also a difference between different types of parking places (e.g. private/ public, free/paid). The number parking places for each home might be more relevant for the origin zone (e.g. homes without a parking place might be less inclined to have and use a car), while public parking places or parking places at work locations might be more relevant at the destination zone (e.g. if there are good parking places at the office, people might be more inclined to take the car to work instead of public transport).

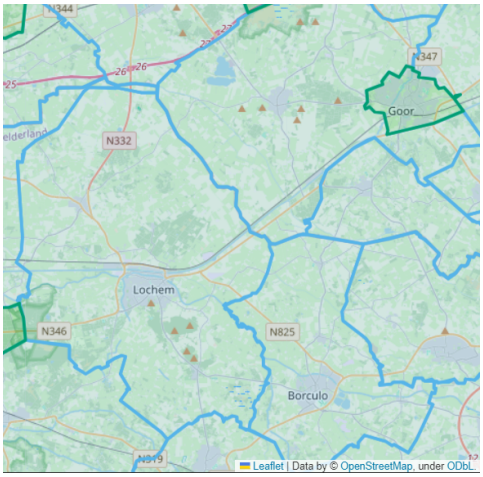

Figure 5.1: The LMS zones for Lochem and Goor. The blue zones indicate the rural clusters and the green zones indicate the small cities & towns cluster. This map was made using the python *Folium* library and RWS WVL (2020)

5.1.5. Limitations of the results

This section discusses the limitations of the results. The previous subsections already covered the limitations with regards to the data, so inaccuracies in the results because of those limitations are not discussed again.

Exploratory data analysis

For the first part of the data analysis, the data was explored to look for interesting trends and initial statistics.

When analysing the modal split of a smaller area (e.g. Amsterdam), there are several zones with only a few trips with a certain mode. When looking at a less dense populated areas, like Zeeland, there are a lot of zones that have no or only a few trips with a certain mode.

For this analysis, all zones with less than 20 departing trips in total were removed from the corresponding figures. However, it is unknown if all zones with more than 20 trips gave a representative impression of that zone. The modal splits according to OViN and LMS were compared and a lot of differences were observed, like how the LMS has more similar travel behaviour over nearby zones. It is possible that the more 'spread out' travel behaviour according to the LMS is a better representation of the real travel behaviour in some cases, compared to OViN that has only a limited number of samples.

Several LMS zones are very large, especially in less dense areas, and consist of several PC4 zones. Maybe there are interesting differences in the modal split within that zone, that are not captured due to the large size of the zones.

Finally, by combining several years of OViN data together, possible trends in travel behaviour that appeared through the years are lost.

Cluster analysis

The cluster analysis was for a large part, a manual process. There was no simple variable that could be used to optimise the process. Because of this, it is not possible to conclude that the 2 final cluster sets are also objectively the best cluster sets. They are 2 cluster sets that satisfy the requirements; are able to capture interesting patterns in the modal split; and are, to the authors knowledge, the best cluster sets that could be made with this set of variables.

The large size of some zones might had an impact on the cluster analysis. An example is given in figure 5.1. This figure shows the cities of Lochem and Goor. Both cities have a similar size and are close to each other, but Goor is part of the small cities & town cluster, while Lochem is placed in the rural cluster of the weighted cluster set. The LMS zone for Lochem includes the large, scarcely populated area around the city, while the LMS zone for Goor only includes the city itself. This affects the D-variables and is likely responsible for putting these zones in separate clusters. In other words, due to the size of the LMS zones, it is possible that some cities were added to a less optimal cluster.

Several assumptions were made to do the PSM. First, the PSM assumes that it is possible to describe different type of persons based on external characteristics (e.g. age, job, type of household) and that those different type of persons have similar preferences with regard to travelling. OViN data did not include a survey or another type of test where people could give their real preferences in travel behaviour. It is still possible that the persons in the different matched clusters have very different preferences in travel behaviour, even though the external demographic characteristics are similar.

Besides that, it is currently assumed that car ownership is a demographic characteristic. However, car ownership is also affected by the spatial environment to some extent (Van Acker et al., 2014; Laviolette et al., 2021). For example, a person that prefers to travel without a car, might buy a car when moving to a more rural area because there are not enough public transport alternatives to travel to their desired activities.

There is a certain randomness in the PSM. Observations in both clusters are matched without replacement. So presumably, some of the people that were not matched, could have been matched if the matching had been tried earlier. In other words, if the matching had been done in another order, the results could have been different. Due to the computational power needed to perform the PSM even once, this was not tested.

5.2. Embedding in current literature

This section compares some of the results that were found with current literature.

5.2.1. Number of clusters and D-variables

When doing the cluster analysis and the PSM, several sources were used. Some studies used the silhouette score to determine the number of clusters and used the value that followed from that (J. Liu et al., 2024; Pot et al., 2023). Other studies used several indices (Patnala et al., 2023; Park et al., 2018) to determine the optimal number of clusters. These methods resulted in 2 to 4 clusters.

This thesis proposes to not simply use the number of clusters that follows from those indices. It can be valuable to use a higher number of clusters. For example, after following the suggestions from the silhouette score for only 2 clusters, the country was divided in 2 regions: large cities and the rest. This gives two regions with vastly different travel behaviour. However, when using too few clusters, smaller areas with deviating travel behaviour (e.g. the medium sized cities) that can provide valuable insights in travel behaviour, are lost.

To conclude, it is important to not only use indicators like the silhouette score when determining the number of clusters, but also the author's expertise.

It is important to note that the above studies often analysed a smaller region, while this thesis analysed a whole country. This can of course affect the number of clusters needed to capture the different regions.

Another noticeable difference with some of the existing literature is the way variables are selected. Many studies used the D-variables to quantify the spatial environment. However, most of these studies did not use all the variables (Kent et al., 2023), and/ or choose only one variable to represent each D-variable, without reflecting on its effectiveness. For example, (J. Liu et al., 2024) used 5 D-variables to cluster the regions, where Distance to transit was only represented by the bus route density or Design only by the intersection density.

This thesis aimed to evaluate as many variables as possible to see which ones were useful for making clusters with different travel behaviour. Instead of just selecting one variable for each D-variable, several were tested to be able to capture the spatial environment as good as possible.

5.2.2. Results propensity score matching

Using PSM, the effect of the spatial environment on travel behaviour is quantified. The results depend a lot on which cluster pair was compared and which mode, but in general the spatial environment is responsible for 50 up to 100% of the differences in travel behaviour between clusters. These percentages are in the same order of magnitudes as other studies.

Patnala et al. (2023) looked at the effect of the spatial environment on the modal split and found that in many cases the spatial environment accounted for more than 50%. For most cluster pairs, the effect

of the spatial environment on active modes was the highest, similar to the results of this thesis. There were however large differences in the effect of the spatial environment between men and women. J. Liu et al. (2024) looked at travel behaviour differences between locals and immigrants for urban and suburban areas. For locals they also found that active mode trips were affected the most by the spatial environment (78%), while car travel and public transport were affected less (62%). This seems to be in line with the thesis, which found very high proportions of the spatial environment for active modes and BTM and lower for train and car travel. So when looking at public transport as a whole (BTM and train) the effect of the spatial environment would presumably be lower and closer to the ATE to OBE ratios for car trips. Park et al. (2018) also found that the spatial environment has a larger effect on walking and public transport trips and a smaller effect on car trips.

Cao et al. (2009) did a literature review on the impact of the spatial environment. Of the 10 studies that tried to quantify the effect of the spatial environment, 8 of those studies indicated that the effect of the spatial environment was stronger than the effect of residential self-selection (ranging from 52 to 90%). However, they also noted that more extensive and complicated methods (and presumably more reliable), made it more difficult to quantify this effect. The conclusion that almost all studies agreed on was that the effect of the spatial environment lessens when the demographic characteristics are taken into account.

This is in line with the findings of this thesis, where in most of the cases the ATE to OBE ratio was less than 1, meaning that the effect of the spatial environment becomes less. There were a lot of differences in the ATE to OBE ratio for the different modes and different cluster pairs, making the answer of what the effect of the spatial environment is not straightforward and dependent on the specific regions that are compared.

All in all, it can be concluded that the findings of this thesis with regard to the effect of the spatial environment are in line with findings from other studies. Even though it is difficult to make any definite conclusions about the exact numbers, the literature and the results find that it is important to take the spatial environment and the demographic characteristics into account when modelling travel behaviour, because both effects are significant.

5.2.3. Regional patterns in travel behaviour

This subsection discusses some of the differences that were seen in the modal split between different clusters and relate this to existing literature.

First of all, cycling patterns. Many studies take cycling and walking together when analysing differences between different regions, including studies done in the Netherlands (e.g. Poorthuis and Zook, 2023; Van De Coevering and Schwanen, 2006). These results then show that the share of cycling increases in more urban areas. This thesis shows that walking and cycling can have vastly different behaviour in different regions in the Netherlands. When looking only at the DU the share of bike trips stays approximately the same, with a small dip at the highest DUs, a trend that is confirmed by Vos (2015). It is possible that cycling forms a competition to public transport in dense areas with a good public transportation network (Kent et al., 2023). Walking trips however, increase significantly with higher population densities. When looking at the different cluster sets, more interesting trends for walking and cycling emerge (e.g. the inner suburbs from the unweighted cluster set, that show the highest share of walking and the lowest share of cycling). By putting walking and cycling in the same category, the implicit assumption is made that both modes show similar trends in the same regions, but this thesis shows that this is not necessarily true. Most studies that reported certain trends in cycling and walking behaviour, are applicable to walking trips in the Netherlands (e.g. increased share of walking with higher road density (Li et al., 2024) or increased share of walking with a better public transport system (Thao & Ohnmacht, 2020)). Different trends are seen for cycling.

Other results from literature are mostly in line with the results of this thesis. The share of public transport trips increases with a higher quality public transport system (Thao & Ohnmacht, 2020); measures like high parking fees are correlated with lower car use (Kent et al., 2023); and a decrease in car trips and increase in public transport trips with higher population density (e.g. Schwanen et al., 2002; Poorthuis and Zook, 2023). The results of this thesis nuances this last part a bit. Sometimes areas with similar population densities show a lot of difference in car use and public transportation use (e.g. the medium-sized cities and the suburbs). The relationships between (population) density and mode use seems to be not as straightforward as often stated in literature.

5.2.4. The LMS documentation and LMS predictions

This subsection discusses the D-variables found in the LMS documentation and compares them with the results of the modal split analysis.

First of all, the unweighted cluster set. This cluster set was based on 7 variables. Interestingly, the variables used for this cluster set are already (indirectly) included in the LMS (see table 4.5 for an overview of the variables of the cluster set and section 3.3.6 for an analysis of the variables in the LMS.). All three Density variables are included in the MDD choice model with one or more variables. The share of service land use variable from the cluster set is presumably correlated with the LMS variables for the number of jobs in the service and retail industry. There are variables in the LMS for parking fee, similarly to the cluster set. The variables road density and the number of tram/metro stops are presumably indirectly included through travel time variables and accessibility constants.

However, the unweighted cluster set is able to uncover interesting trends in travel behaviour that the LMS is not able to capture (e.g. the medium sized cities). This observation implies that it might be possible to improve the LMS by simply implementing already existing variables in a better way.

For the weighted cluster set, more different variables were used that are currently not used in the LMS, e.g. share of houses built before 1945 or distance to points of interest.

Most spatial environment variables in the LMS are related to the destination zone (i.e. the destination of the first trip of a tour) and some are related to the origin. The modal split analysis focused on the characteristics of the origin zone. However, trips for the whole day are included (and not only the morning peak, for example). This means that in approximately half of the trips, the origin zone in the analysis is the 'destination' zone, when considering tours.

In section 3.4 it was found that the fewest D-variables are assigned to BTM, cycling and walking. Based on the modal split analysis, it was found that the LMS models BTM relatively accurate, especially for the areas with a larger BTM network. Walking trips were underestimated, but the trends were modelled well. Based on this, it can be concluded that the LMS is able to model the share of these two modes for different regions relatively well. However, there are presumably additional variables needed to model bike trips.

5.3. Generalisability

This thesis researched differences in modal split between different regions in the Netherlands, with a focus on how these differences can be implemented in transport models.

The exact differences that were observed between OViN and LMS and the specific advises following those differences are presumably not directly applicable to other transport models (e.g. walking is systematically underestimated). However, the differences found in the LMS can still provide insights for other transport models. For example, it was observed that car trips were highly overestimated in medium-sized cities. This is likely partly because of the DU (the unweighted cluster set showed that clusters with a similar DU were predicted to have similar levels of car travel). If other transport models also rely on a variable like the DU, it might be valuable to check if similar differences can be found between the 'real' and predicted travel behaviour. This is of course no guarantee. Other models could have inaccuracies on very different aspects.

The clustering method itself could be applied to other models. Data for the D-variables can be gathered for the specific area of the model and the zones can be clustered. It is difficult to say if using the same variables will give similar clusters in a different region. It might be possible that some of the variables used in these cluster sets were only proxies for other variables (e.g. Ewing and Cervero (2010) suggested that maybe Density is more of an intermediate variable and could also be expressed by other D-variables). There is no guarantee that variables have the same effect in different areas, especially when looking at different countries. Differences in modal split between countries might not only be because of different demographics or different spatial environment characteristics, but differences in culture and policies can also have a large effect. One of the reasons for a high share of cycling is that the Dutch government devoted a lot of energy to promoting cycling and one of the reasons for high public transportation use is a policy that resulted in an increase in offices close to train stations (Vos, 2015).

To conclude, the exact results found in this thesis might not be directly applicable to other countries or other models than the LMS. However, the insights obtained from this thesis can help to better understand differences in travel behaviour between regions and can help identifying specific areas that show irregular trends in travel behaviour that will need extra attention in modelling or research. Besides that, the methodologies used (hierarchical clustering of D-variables and PSM) can also be applied to other areas.

6

Conclusions and recommendations

6.1. Conclusions

This section first answers the sub-questions based on the literature review and the modal split analysis. After that, the main research question is answered.

6.1.1. Answers sub-questions

What region specific factors affect travel behaviour?

This sub-question can be answered based on the literature review.

A common framework for quantifying the spatial environment are the D-variables, as described by Ewing and Cervero (2010). The 6 D-variables: Density, Diversity, Design, Destination accessibility, Distance to transit and Demand management, can help getting a comprehensive overview and quantifying the spatial environment. It is also important to include the Demographics. Even though it is not a characteristic from the spatial environment, it is still an important factor to control for. The D-variables themselves are still abstract, but they can be expressed with other variables that can be measured (e.g. Density could be expressed as the population density and the job density).

The D-variables are not independent variables, but they are correlated. This means that it is not always possible to determine the effect of a specific D-variable, because the effect can be caused by a combination of variables (Dieleman et al., 2002).

In which ways are different travel behaviours in different regions captured in the Dutch national transport model?

The LMS is an extensive transport model that takes many different variables into account, including variables related to the spatial environment. It often use the degree of urbanisation to divide zones and model travel behaviour, which is based on the population density. To calculate the degree of urbanisation, not only the population density of a zone itself is used, but also the population density of surrounding zones.

This section summarizes the factors related to the spatial environment in the LMS documentation, and relates them to the D-variables from the previous section.

The LMS consists of several modules. First there is the population module. The degree of urbanisation is used when calibrating the household distributions for each zone. Secondly, there is the car ownership module. In this module, several Density and Demand management variables are used, together with one Diversity variable. For determining the travel frequency, mainly Density variables are used combined with a variable for the accessibility, which is presumably a combination of different D-variables. The final part of the LMS that is relevant for this thesis, determines the mode, destination and part of day for each trip, using a large nested logit model with many variables. Most of the spatial environment variables in this module are Density or Distance to transit variables. The Distance to transit variables are almost exclusively for modelling train travel. There are also several variables that are

related to the other four D-variables.

To conclude, even though most D-variables are implemented in the LMS to some extent, their representation is not equal. The majority of the variables are related to Density or Distance to transit. When looking at the degree of urbanisation variables, several degrees are often grouped together in one variable, which gives less differentiation between different regions.

How does actual and predicted travel behaviour differ between regions with a different degree of urbanisation?

Throughout the country large differences in travel behaviour can be observed, for both OViN (actual travel behaviour) and the LMS (predicted travel behaviour). Zones with a degree of urbanisation of 6 are characterized by low levels of car travel and a high shares of public transport and walking. For each increase in the degree of urbanisation, the amount of car travel decreases and the shares of public transport and walking increase. Bike use stays relatively the same for all degrees of urbanisation.

The LMS is able to model these trends relatively well, although car travel appears to be systematically overestimated and walking underestimated. Based on these aggregated results, both the choice for using the degree of urbanisation as a variable in the LMS and the predicted modal split following from the LMS seem logical. In other words, when comparing regions with a different degree of urbanisation on national level, the predictions from the LMS are in line with the observations in OViN. This is a logical conclusion, because OViN data has been used to calibrate the LMS.

How does actual and predicted travel behaviour differ between regions with a similar degree of urbanisation and what could be the cause of those potential differences?

When focusing on how travel behaviour differs within similar degrees of urbanisation, more interesting trends can be observed in OViN that are not always captured by the LMS. By looking at smaller regions (e.g. the municipality of Amsterdam) it was observed that there are large differences in modal split between zones within the same degree of urbanisation. The LMS, however, predicts similar levels of mode use for neighbouring zones, losing some of the trends that can be observed in OViN. This effect is the strongest when looking at the car driver trips and bike trips.

In the cluster analysis, zones were clustered together based on the D-variables that were identified in the literature review. Two cluster sets were made, that both have seven clusters: the weighted and unweighted cluster set. These cluster sets contain clusters like the centres of large urban areas, centres of medium-sized cities, suburbs, small cities and rural areas. Rural areas and smaller cities are characterized by high car use and low shares of public transport and walking, while centres of large urban areas are characterized by the opposite. Medium-sized city centres have low use of car and BTM, high shares of train and bike use and average shares of walking. In contrast, suburbs of large urban areas have high shares of BTM and walking, low shares of car and bike and average share of train use.

The cluster analysis showed that clusters with a similar degree of urbanisation, could have a very different modal split. Especially for the share of car drivers, the LMS seemed to predict similar shares of trips for clusters with the same degree of urbanisation. For bike, the LMS predicts similar levels of use for all zones with only small deviations, regardless of the degree of urbanisation. OViN however shows that bike use can differ a lot in different zones.

These findings imply that actual travel behaviour of zones with the same degree of urbanisation can have large differences, but the LMS lacks the ability to accurately differentiate between zones that have different spatial environment characteristics and the same degree of urbanisation. For example, the centres and suburbs of the medium-sized cities from the unweighted cluster set (cluster 2 & 3) have a similar degree of urbanisation, but the centre cluster has a significantly higher share of service land use and a higher parking fee. The differences in car driver use and bike use between these clusters are large, even after controlling for the demography. This is not captured by the LMS.

This is in line with the hypothesis that the degree of urbanisation alone is not enough to model travel behaviour in different regions. In other words, while the degree of urbanisation might be a suitable variable in transport models to capture trends in travel behaviour on an aggregated scale, the degree of urbanisation alone does not contain enough information about the spatial environment to accurately

capture the real modal split for car drivers and bike use.

Different trends between regions for the other modes (car passenger, train, BTM and walking) are more accurately captured by the LMS, even though the absolute values might differ a lot (walking). These modes are presumably easier to model using the degree of urbanisation.

How could the Dutch national transport model be improved to capture the differences in travel behaviour in different regions more realistically?

Sikder et al. (2013) suggested that it might be valuable to investigate whether it is possible to identify different type of regions that are able to capture differences in travel behaviour between different regions. Such different types of regions were created using a hierarchical cluster technique in combination with the D-variables found in the literature.

From these clusters, regions with interesting travel patterns emerged that are not identified when using only the degree of urbanisation. Those regions showed large differences in modal split between OViN and LMS, suggesting that the LMS is not capable in capturing those regions. Propensity score matching showed that the differences between regions are not only caused by demographic characteristics, but that the spatial environment plays an important role. These findings underline the importance of including spatial environment variables in a transport model, because when including only demographic characteristics not all travel behaviour can be captured. When including the spatial environment variables in a transport model, it is important to not only include the degree of urbanisation or other Density-related variables. Other D-variables like Diversity (share of service land use or the share of older houses) or Demand management (parking fee) have shown to play an important part in creating clusters with different travel behaviour. This includes patterns in travel behaviour that would not have been discovered when looking only at Density variables.

Further research is needed to show to what extent the LMS can be improved by replacing the degree of urbanisation with those new clusters. It was not possible to test this in this thesis. However, this thesis shows that using only the degree of urbanisation to quantify the spatial environment is not enough. It is possible that these clusters can provide a way to improve the LMS, but this is not the only way a transport model can be improved. The individual D-variables can also be implemented as (dummy) variables in a transport model, similarly to how the LMS currently uses variables related to the spatial environment. The clusters are not needed for this.

6.1.2. Answer main research question

Now all the sub-questions are answered, an answer to the main question can be formulated:

> *To what extent does the degree of urbanisation capture the difference in travel behaviour in different regions in current transport models and in what ways can these differences be captured more realistically with those same transport models?*

The degree of urbanisation is a variable based on the population density. The data analysis showed that by looking only at the population density, important nuances are lost. It is true that the degree of urbanisation shows clear and predictable differences in travel behaviour, although the differences in modal split between zones with a degree of 1 and 2 are not significant. All Density related variables show a strong correlation with most of the other D-variables and with the different modes, making the degree of urbanisation a logical choice as a variable to represent the spatial environment. However, there are regions with similar population densities, but with a very different modal split. The most interesting results from the cluster analysis were the medium-sized city centres and the suburbs of the large urban areas. The medium-sized cities had a lower population density than the suburbs, showed similar levels of car use, but significantly higher train and bike use, while the suburbs showed high shares of walking and BTM. This is an example of nuances that are lost when looking only at the degree of urbanisation, which shows only increasing levels of train, BTM and walking with a higher population density and decreasing levels of car use.

The LMS is able to capture the trends in different regions for car passenger, train, BTM and walking relatively well, although the share of walking is significantly underestimated. The LMS seems to be better in capturing a larger number of trips, because the relative errors grow when modelling a lower number of trips. This can partly be explained by the fact that OViN is less reliable for a lower number of observations.

The different trends shown by bike and car use are often not captured well by the LMS. Based on the results from the cluster analysis, the number of car trips that are modelled per cluster seems to be heavily affected by the degree of urbanisation. This gives large errors in clusters that do not follow the trend of the degree of urbanisation (e.g. the medium sized cities). Besides that, the absolute and relative errors in the number of car trips seem to be higher in clusters with lower car travel. For bike use, the errors are less correlated with the degree of urbanisation, but are presumably caused by an (implicit) assumption that bike use differs relatively little throughout the country. This assumption is true when looking at the degree of urbanisation, but not true when looking at the different clusters. The exploratory analysis showed that large differences in bike use are observed in neighbouring zones. This is true for both urban and rural areas.

The differences in modal split between the different degrees of urbanisation and the different clusters are caused by both demographic characteristics and differences in the spatial environment, as shown by the propensity score matching. On average, differences in BTM, bike and walking are for a very large part dependent on differences in the spatial environment (more than 80% on average). Differences in car and train use are for a smaller part dependent on the spatial environment, though the effect is still more than 50% on average. This is with the side note, that there are several cluster pairs where insignificant differences were found after matching, meaning that the effect of the spatial environment is presumably very little to none, which would lower the average. This shows that the effect of the spatial environment is highly dependent on the regions that are compared.

To conclude, characteristics of the spatial environment play a large roll in the differences in modal split that can be observed between different regions. In general, car driver and bike use will need the most improvement in capturing the right trends. The trends in walking are captured relatively well, although the absolute number of trips is underestimated. BTM and train are both captured accurately, although train travel is modelled more accurately on zone-level when zooming in. The biggest difference between those two modes is that BTM use is affected by the spatial environment significantly more, than train use. This means that modelling BTM might benefit more from additional D-variables. Car passenger use is also captured relatively well by the LMS, without large irregularities.

When improving the modal split of the LMS, it is important to keep in mind the above mentioned points. The results showed that different cluster sets are able to uncover previously hidden trends in travel behaviour. These, or other, cluster sets could be implemented in the LMS in a way similar to how the degree of urbanisation is now used. Further research is needed to examine whether this improves the ability of the LMS to distinguish between different regions. Another possibility would be to not replace the degree of urbanisation, but to add a few extra dummy variables for certain areas that have large errors in the LMS, that cannot be easily solved by using the degree of urbanisation. It is further possible to make those dummy variables only applicable to certain modes. For example, some clusters showed no significant differences in the share of one mode, but did show significant differences in the share of other modes. This means that a cluster does not have to be relevant for each single aspect of travel behaviour to be an improvement to the model. Finally, this thesis showed that the characteristics of the spatial environment can be expressed using the D-variables. In the cluster analysis, it was shown that the D-variables vary a lot over different regions. By implementing a larger variety of D-variables in the nested logit model in the LMS, it is possible that differences in the LMS can be captured more realistically.

The above mentioned ways to capture regional differences in transport models are still very focused on the LMS. The exact differences that were found between OViN and LMS might not be relevant for other transport models. However, the methods and recommendations mentioned above can still be relevant in other contexts and for other transport models.

6.2. Recommendations

This section gives recommendations based on the results of this thesis. The first part of this section reflects on how these results affect current users of the LMS, like ProRail. This subsection includes some recommendations on how the current version of the LMS should be used. Next, recommendations are given for future scientific research and for more practical applications.

6.2.1. Consequences for policy makers

The outcomes of the LMS are currently used by several organisations like the IenW and ProRail. The LMS plays an important role in policy making (Hofman, 2017). This thesis sheds more light on inaccuracies in the LMS with regard to the modal split, which might have consequences for the reliability of the LMS results.

First of all, it is important to explain which output of the LMS was analysed in this thesis. The LMS results used in this thesis are the synthetic matrices from the base year 2018, which directly follow from the mode-destination-part of the day nested logit model. The transport forecast for future years is obtained by multiplying the growth factor by transport in the base year[1]. This base year gives an estimation of the traffic for each zone in 2018, based on extensive data, and is meant to be a realistic reflection of all traffic in the Netherlands. The growth factor is obtained by dividing the synthetic forecast year matrix by the synthetic base year matrix, so the synthetic matrices are only used for the relative growth in traffic (RWS WVL, 2021j; RWS WVL, 2021b). In other words, the synthetic matrices of the LMS are not directly used for the transport forecasts and in policy making, but are first combined with more accurate data. The non-synthetic base year and forecast matrices were not analysed in this thesis.

Because of the use of the base year, small inaccuracies in the synthetic matrices of the LMS should not automatically lead to large errors in the predictions for future years. This does not mean that the inaccuracies do not matter. Wrong predictions in mode use in the synthetic matrices will lead to inaccurate growth factors. This will give inaccurate predictions, even if the non-synthetic base year is accurate.

It is possible to run scenarios in the LMS, based on policies that need to be tested (Hofman, 2017). The more the synthetic matrices differ from reality, the less suitable the model will be to explain differences in travel behaviour due to policy changes. This is illustrated with two examples. The LMS overestimates car driver use, especially in city centres. If those city centres want to implement policies to reduce car use, the LMS might be less suitable to model the consequences of these policies in the future. The LMS is currently unable to accurately capture car use in those locations, presumably due to the lack of variables that effectively capture the differences in the spatial environment or due to incorrect modelling of the population distribution in each zone. If the true effects of already existing variables are unknown, it is even more difficult to determine whether the LMS is able to capture the effect of the new policies on the modal split in a realistic way, especially if those policies are related to the spatial environment. In other words, implementing policies or future scenarios in the LMS that change aspects of the spatial environment, might give additional uncertainty in the forecasts of certain regions. This is something the users of the LMS should be aware of.

Another example would be policies aimed to stimulate bike use. The LMS predictions for bike use are barely affected by the spatial environment. However, this way policies might overlook the large effect the spatial environment has on bike use in reality. Policy makers might be able to use these effects of the spatial environment to their advantage, if they know of their existence. Besides that, it is possible that some policies might be more effective in one region, compared to another region due to differences in the spatial environment. This will not be visible as result if this policy is simulated with the LMS.

When keeping these shortcomings in mind, the LMS is still a useful tool to predict future transport. Users of the LMS should not follow the forecasts blindly, but critically evaluate the results and compare them with trends that can currently be observed in different regions.

For ProRail specifically, train use seems to be modelled relatively well when comparing the average values from the clusters. When zooming in, the LMS was often able to correctly identify zones with a high or low train use, although the exact predictions often differed from OViN. The (relative) errors seemed to increase in more rural areas. For ProRail this means that the absolute number of train trips on an aggregated level is fairly accurate. On a smaller scale, the number of trips will become less accurate, but zones with high and low use are often correctly identified. This will presumably be favourable for determining the growth factors.

Cellissen et al. (2022) identified that the LMS over- or underestimates the growth of certain train stations and train lines. It is unknown if these differences are due to the inaccuracies in the number of train trips or because of inaccuracies in modelling the station choice or destination choice. To identify

[1] The exact method to obtain the forecast matrices also accounts for special cases.

this, additional research is needed.

To conclude, this thesis argues that transport models can be improved by implementing additional variables in the model that are related to the spatial environment (the D-variables). By adding more information to the model, the LMS could give more reliable predictions which can lead to better justifications for certain policies. Almost all data used to gather the variables came from the CBS or was already available in the LMS. In other words, it is not needed to gather a lot of new data to improve the LMS. Most data is already (publicly) available.

6.2.2. Scientific recommendations

First, as explained in section 5.1.1, this thesis focused primarily on the modal split and not on other aspects of travel behaviour like distance, duration, departure time and travel motives. Future research should study the effect of the different cluster sets on those other aspects of travel behaviour. This is needed before more definite conclusions can be made about the quality of the different cluster sets. For example, the cluster pairs 3 & 6 (the suburbs of the medium-sized cities and the towns & small cities) and 5 & 2 (the suburbs from the large urban areas and the centres of medium-sized cities) from the weighted cluster set show parallels in their differences in travel behaviour, as described in section 4.3.2. It would be interesting to see if those same parallels can also be seen in other aspects of travel behaviour (e.g. cluster 3 and 5 are more BTM and walking oriented, while cluster 6 and 2 are more train and bike oriented. It can be researched how this affects their travel distances or how these differences in mode uses are distributed over the different travel motives).

Another difference between OViN and LMS that could maybe be further explained by looking at travel distance, is the overestimation of car use and underestimation of walking. It is possible that the LMS models a lot of (short) trips by car, which are in reality done by walking. However, it could also mean that the LMS models more long-distances trips by car instead of shorter distance trips by foot. In the last case, there should be a substantial difference between the average travel distance modelled by the LMS and the distance observed in OViN.

In this thesis, only characteristics of the origin zone were included. For future research, it would be good to also include characteristics of the destination zone, because this is relevant to the mode choice. Another possibility would be to include the ‘home zone’ of each person. For example, it might be possible that a person living in a rural area and a person (with similar demographic characteristics) living in the centre of a large urban area will choose different modes for the same trip, because of their different experiences with each mode.

It could also be beneficial to include variables from the zones around the origin and destination, instead of only the origin and destination itself. The cluster analysis already showed how there could be large differences between the population density of a zone and the population density that included surrounding zones. The effect of similar ‘surrounding’-variables could be researched.

Secondly, propensity score matching was done on the whole dataset and the effect of the spatial environment was estimated for the whole population. Patnala et al. (2023) found that there can be large differences between how men and women are affected by the spatial environment and J. Liu et al. (2024) found that the spatial environment affects the travel behaviour of immigrants and locals in a different way. For future research, it would be interesting to look at how different population segments within the clusters are affected by the spatial environment differently. This does not only improve the understanding of how people are affected by travel behaviour, but it can also be utilized in transport models if significant differences are found. For example, dummy variables for the spatial environment can be added that are only applicable to women or people from a certain age group. This way the effects of the spatial environment can be modeled in more detail.

Similar variables are already implemented in the car ownership module: There are two variables for disposable income when the household has a certain degree of urbanisation (RWS WVL, 2021d). It might be beneficial to research whether the LMS (or other transport models) can be improved by adding more of these ‘combined’ variables.

Thirdly, this thesis focused on the LMS which is a transport model for the whole country. It would be interesting to research what the effects are when a similar methodology is applied to a smaller region.

For this research, it was chosen to limit the number of clusters to prevent creating complicated cluster sets and to prevent overfitting. When using a smaller region with smaller zones, a similar number of clusters can show more details and nuances, than 7 clusters divided over the whole country. The unweighted cluster set for this thesis already showed that by dividing the zones with a degree of urbanisation of 6 into two clusters, interesting differences emerged. So even a smaller region can show differences in travel behaviour. A more practical example where this could be applied is the NRM, which is a model related to the LMS but focused on only a part of the country.

Fourthly, the preliminary literature review that was done before the start of this thesis (see appendix A) found that a certain decision making theory (e.g. utility optimization), might be the best theory to use for one region, but less than optimal for another region. Similarly, the choice of a certain model (e.g. nested logit) also brings in additional implicit assumptions that might not be true for each region (Sikder et al., 2013). This means that the current model structure in the LMS might not be the optimal logit structure for each region. This idea is both theoretically and practically difficult to research and implement and falls outside the scope of this thesis. However, this could be further researched in the future. If there is a wrong assumption somewhere at the foundation of a model, it would presumably be difficult to correct this on higher levels of the model.

Finally, this thesis focused on how the degree of urbanisation affects mode choice. However, the degree of urbanisation is also used on a deeper level in the LMS: the distribution of the population and modelling car ownership. These factors can also affect how transport is modelled. It is possible that differences that were seen between the LMS and OViN have less to do with the variables used in the mode-destination-part of day choice model, and more with other factors like the car ownership model. The share of car drivers was often overestimated, though this mode was only ‘available’ when the type of person is in possession of a car and has a driver’s licence (RWS WVL, 2021f). It would be good to further research which (presumably Density-related variables) are responsible for this overestimation of the share of car drivers.

Besides that, propensity score matching showed that both the spatial environment and residential self-selection play a significant part in explaining the differences in modal split between different regions. If the household distribution modelled by the LMS gives an inaccurate representation of the real household distribution of a zone, it is possible that (the lack of) differences in modal split between zones are caused by the wrong population. For future research, it would be good to investigate which differences in modal split between OViN and LMS are caused by the lack of D-variables included in the LMS and which differences are caused by an inaccurate household distribution.

6.2.3. Practical recommendations

Some of the practical recommendations of how the LMS can be improved followed directly from the answer of the main research question and will not be repeated again. Other recommendations followed from additional scientific research and are stated in the previous section. This section will provide some extra recommendations, that are more focused on the process or possible ideas that were not researched in this thesis.

First of all, the limited size of the data causes less reliable results. The LMS is calibrated using 3 stacked years of OViN data. The total number of trips in OViN seems high, but when spreading those trips out over all LMS zones, it gets significantly lower. There are zones that contain only a few data points or no data points for some of the less popular modes. A straightforward way to improve the LMS is to collect more data for calibration. Future research could analyse the results of using more than 3 years of OViN data.

The discussion section already stated that a lot of assumptions and simplifications had to be made when collecting and processing both the OViN data and the data for the D-variables. Before any of the suggestions here can be implemented, it is important to gather more exact data, such that the quality of the data is on a similar level as the data that is already used in the LMS. Besides that, wrongly assigned zones should be identified and removed from the clusters.

Creating the clusters was done manually in this thesis, due to the lack of good indicators to optimize.

It might be interesting to find indicators that can be used to automate this process.

As mentioned before in this thesis, the LMS seems to ‘spread out’ travel behaviour over neighbouring zones. This can be a good thing in some cases, because outliers observed in OViN are removed, due to the lack of data. In some cases, however, a peak in mode use in OViN can also be seen in the LMS. For example, in The Hague there is one zone that shows very high levels in car passenger use compared to the surrounding zones, according to the LMS (see figure 4.8). This peak zone can also be seen in OViN. It is not clear if there is indeed a unusual large peak in car passenger use in that zone, or if the LMS wrongly overestimated car use in that zone due to an outlier in the OViN data. A similar example can be observed for train use in Amsterdam (see figure 4.6).

So, the LMS should be checked for overfitting in these zones for these specific modes, whether the observed effects can also be seen in reality.

Bibliography

Berman, J. J. (2016, January). Understanding your data. In *Elsevier ebooks* (pp. 135–187). https://doi.org/10.1016/b978-0-12-803781-2.00004-7

Bovy, P., Bliemer, M., & Van Nes, R. (2006). Course ct4801: Transportation modeling. *Delft: Delft University of Technology, Faculty of Civil Engineering and Geosciences, Transport & Planning Section*.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in statistics. Theory and methods/Communications in statistics, theory and methods*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Cao, X. (2014). Residential self-selection in the relationships between the built environment and travel behavior: Introduction to the special issue. *Journal of Transport and Land Use*, *7*(3), 1–3.

Cao, X., Mokhtarian, P. L., & Handy, S. (2009). Examining the Impacts of Residential Self□Selection on Travel Behaviour: A Focus on Empirical Findings. *Transport Reviews*, *29*(3), 359–395. https://doi.org/10.1080/01441640802539195

Cao, X., Xu, Z., & Fan, Y. (2010). Exploring the connections among residential location, self-selection, and driving: Propensity score matching with multiple treatments. *Transportation research. Part A, Policy and practice*, *44*(10), 797–805. https://doi.org/10.1016/j.tra.2010.07.010

Castiglione, J., Bradley, M., & Gliebe, J. (2014, March). *Activity-Based Travel Demand Models: A primer*. https://doi.org/10.17226/22357

Cellissen, R., Kouwenhoven, M., & Hogenberg, J. G. (2022, October). *Backcast LMS: Vergelijking van prognoses en waargenomen ontwikkelingen* (tech. rep.). Colloquium Vervoersplanologisch Speurwerk. https://www.cvs-congres.nl/e2/site/cvs/custom/site/upload/file/paper_search/2022/cvs_127_backcast_lms_vergelijking_van_prognoses_en_waargenomen_ontwikkelingen_1_2022%20(002).pdf

Centraal Bureau voor de Statistiek. (n.d.). Stedelijkheid (van een gebied). https://www.cbs.nl/nl-nl/onze-diensten/methoden/begrippen/stedelijkheid--van-een-gebied--

Centraal Bureau voor de Statistiek. (2012, April 4). Nabijheidsstatistiek: hoe ver wonen Nederlanders van voorzieningen? Retrieved July 22, 2024, from https://www.cbs.nl/nl-nl/achtergrond/2012/14/nabijheidsstatistiek-hoe-ver-wonen-nederlanders-van-voorzieningen-

Centraal Bureau voor de Statistiek. (2018, December). *Onderweg in Nederland (ODiN) 2018* (tech. rep.).

Centraal Bureau voor de Statistiek. (2022a). *Bodemgebruik, wijk- en buurtcijfers 2017*. https://opendata.cbs.nl/portal.html?_la=nl&_catalog=CBS&tableId=85217NED&_theme=306

Centraal Bureau voor de Statistiek. (2022b, March). *Bestand Bodemgebruik 2017 (BBG2017): Productbeschrijving (v2)* (tech. rep.).

Centraal Bureau voor de Statistiek. (2023, June). *Onderweg in Nederland (ODiN) 2022* (tech. rep.).

Centraal Bureau voor de Statistiek & ESRI Nederland. (2019). *Statistische gegevens per vierkant en postcode 2019*. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode

Centraal Bureau voor de Statistiek & Kadaster. (2019). *Wijk- en Buurtkaart 2017*. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2017

Centraal Bureau voor de Statistiek & Rijkswaterstaat. (2014). *Onderzoek Verplaatsingen in Nederland 2013 - OViN 2013*. https://doi.org/10.17026/dans-x9h-dsdg

Centraal Bureau voor de Statistiek & Rijkswaterstaat. (2015). *Onderzoek Verplaatsingen in Nederland 2014 - OViN 2014*. https://doi.org/10.17026/dans-x95-5p7y

Centraal Bureau voor de Statistiek & Rijkswaterstaat. (2017a). *Onderzoek Verplaatsingen in Nederland 2015 - OViN 2015 versie 2.0*. https://doi.org/10.17026/dans-z2v-c39p

Centraal Bureau voor de Statistiek & Rijkswaterstaat. (2017b). *Onderzoek Verplaatsingen in Nederland 2016 - OViN 2016*. https://doi.org/10.17026/dans-293-wvf7

Centraal Bureau voor de Statistiek & Rijkswaterstaat. (2017c). *Onderzoek Verplaatsingen in Nederland 2017 - OViN 2017*. https://doi.org/10.17026/dans-xxt-9d28

Cervero, R. (2002). Built environments and mode choice: toward a normative framework. *Transportation Research Part D: Transport and Environment*, *7*(4), 265–284. https://doi.org/10.1016/s1361-9209(01)00024-4

Dargay, J., & Hanly, M. (2003). The impact of land use patterns on travel behaviour. https://www.researchgate.net/profile/Joyce-Dargay/publication/228915810_The_Impact_of_land_use_patterns_on_travel_behaviour/links/0deec5331bfb31eeaf000000/The-Impact-of-land-use-patterns-on-travel-behaviour.pdf

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/tpami.1979.4766909

de Jong, G., Tuinenga, J. G., & Kouwenhoven, M. (2008, November). *Prognoses van het Landelijk Model Systeem: komen ze uit?* (Tech. rep.). Colloquium Vervoersplanologisch Speurwerk. https://www.cvs-congres.nl/cvspdfdocs/cvs08_62.pdf

de Souto, M. C. P., de Araujo, D. S. A., Costa, I. G., Soares, R. G. F., Ludermir, T. B., & Schliep, A. (2008). Comparative study on normalization procedures for cluster analysis of gene expression datasets. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2792–2798. https://doi.org/10.1109/IJCNN.2008.4634191

Dieleman, F. M., Dijst, M., & Burghouwt, G. (2002). Urban Form and Travel Behaviour: Micro-level Household Attributes and Residential Context. *Urban Studies*, *39*(3), 507–527. https://doi.org/10.1080/00420980220112801

Ettema, D., & Nieuwenhuis, R. (2017). Residential self-selection and travel behaviour: What are the effects of attitudes, reasons for location choice and the built environment? *Journal of transport geography*, *59*, 146–155. https://doi.org/10.1016/j.jtrangeo.2017.01.009

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

Ewing, R., & Hamidi, S. (2015). Compactness versus Sprawl. *Journal of Planning Literature*, *30*(4), 413–432. https://doi.org/10.1177/0885412215595439

Feng, J., Dijst, M., Prillwitz, J., & Wissink, B. (2013). Travel Time and Distance in International Perspective: A Comparison between Nanjing (China) and the Randstad (The Netherlands). *Urban Studies*, *50*(14), 2993–3010. https://doi.org/10.1177/0042098013482504

Harts, J., Maat, C., & Van Emmichoven D, Z. (1999). Meervoudig stedelijk ruimtegebruik; methode en analyse. http://resolver.tudelft.nl/uuid:25561384-f74e-4319-85ae-3dc57d6f3d24

Hofman, F. (2017, December). *Het Landelijk Model Systeem* (tech. rep. No. 763671). Ministerie van Infrastructuur en Waterstaat, Rijkswaterstaat Water, Verkeer en Leefomgeving (RWS, WVL). https://open.rijkswaterstaat.nl/publish/pages/34566/verkeer_en_vervoer_het_landelijk_model_sysyteem.pdf

Hong, J., Shen, Q., & Zhang, L. (2013). How do built-environment factors affect travel behavior? A spatial analysis at different geographic scales. *Transportation*, *41*(3), 419–440. https://doi.org/10.1007/s11116-013-9462-9

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Unsupervised learning. In *An introduction to statistical learning: With applications in python*. Springer.

Kemperman, A., & Timmermans, H. H. (2012). Environmental Correlates of Active Travel Behavior of Children. *Environment and Behavior*, *46*(5), 583–608. https://doi.org/10.1177/0013916512466662

Kent, J., Crane, M., Waidyatillake, N. T., Stevenson, M., & Pearson, L. (2023). Urban form and physical activity through transport: a review based on the d-variable framework. *Transport reviews*, *43*(4), 726–754. https://doi.org/10.1080/01441647.2023.2165575

Kockelman, K. M. (1997). Travel Behavior as Function of Accessibility, Land Use Mixing, and Land Use Balance: Evidence from San Francisco Bay Area. *Transportation research record*, *1607*(1), 116–125. https://doi.org/10.3141/1607-16

Laviolette, J., Morency, C., Waygood, O. D., & Goulias, K. G. (2021). Car Ownership and the built environment: a spatial modeling approach. *Transportation research record*, *2676*(3), 125–141. https://doi.org/10.1177/03611981211049409

Li, X., Wang, Z., Yu, L., & Xie, B. (2024). Exploring the gap in people's travel behavior between urban villages and commercial housing: The role of built environment. *Travel behaviour and society/-Travel behaviour society*, *36*, 100794. https://doi.org/10.1016/j.tbs.2024.100794

Limtanakool, N., Dijst, M., & Schwanen, T. (2006). The influence of socioeconomic characteristics, land use and travel time considerations on mode choice for medium- and longer-distance trips. *Journal of Transport Geography*, *14*(5), 327–341. https://doi.org/10.1016/j.jtrangeo.2005.06.004

Linh, H. T., Adnan, M., Ectors, W., Kochan, B., Bellemans, T., & Tuan, V. A. (2019). Exploring the spatial transferability of feathers – an activity based travel demand model – for ho chi minh city, vietnam [The 10th International Conference on Ambient Systems, Networks and Technologies (ANT 2019) / The 2nd International Conference on Emerging Data and Industry 4.0 (EDI40 2019) / Affiliated Workshops]. *Procedia Computer Science*, *151*, 226–233. https://doi.org/10.1016/j.procs.2019.04.033

Liu, J., Xiao, L.-N., & Wang, B. (2024). The varying effects of residential built environment on travel behavior of internal migrants and locals. *Travel behaviour and society/Travel behaviour society*, *34*, 100692. https://doi.org/10.1016/j.tbs.2023.100692

Liu, L., Wang, H., & Duan, J. (2023). How streetscape affects car use: Examining unexamined features of urban environment with fine-grained data. *Cities*, *132*, 104096. https://doi.org/10.1016/j.cities.2022.104096

Liu, T., & Ding, C. (2024). Revisiting built environment and travel behavior: A natural experiment accounting for residential self-selection. *Journal of transport geography*, *115*, 103794. https://doi.org/10.1016/j.jtrangeo.2024.103794

McArthur, D. P., Kleppe, G., Thorsen, I., & Ubøe, J. (2011). The spatial transferability of parameters in a gravity model of commuting flows. *Journal of Transport Geography*, *19*(4), 596–605. https://doi.org/10.1016/j.jtrangeo.2010.06.014

Millward, H., & Spinney, J. (2011). Time use, travel behavior, and the rural–urban continuum: results from the Halifax STAR project. *Journal of Transport Geography*, *19*(1), 51–58. https://doi.org/10.1016/j.jtrangeo.2009.12.005

Ministerie van Infrastructuur en Waterstaat. (2023, November). Verkeers- en vervoermodellen LMS en NRM. Retrieved January 4, 2024, from https://www.rijkswaterstaat.nl/zakelijk/open-data/modellen-en-applicaties/verkeers-en-vervoermodellen/jaarlijkse-prognoses/prognoses-maken/lms-en-nrm

Mokhtarian, P. L., & Van Herick, D. (2016). Viewpoint: Quantifying residential self-selection effects: A review of methods and findings from applications of propensity score and sample selection approaches. *Journal of transport and land use*, *9*(1). https://doi.org/10.5198/jtlu.2016.788

Næss, P. (2006). Accessibility, Activity Participation and Location of Activities: Exploring the Links between Residential Location and Travel Behaviour. *Urban Studies*, *43*(3), 627–652. https://doi.org/10.1080/00420980500534677

Næss, P., Cao, J., & Strand, A. (2017). Which D's are the important ones? The effects of regional location and density on driving distance in Oslo and Stavanger. *Journal of Transport and Land Use*, *10*(1). https://doi.org/10.5198/jtlu.2017.1183

Nationaal Wegenbestand. (2021). 30.000 km fietspad toegevoegd aan het NWB. Retrieved June 10, 2024, from https://www.nationaalwegenbestand.nl/nieuws/30000-km-fietspad-toegevoegd-aan-het-nwb

Oakes, J. M., & Johnson, P. J. (2006). Propensity score matching for social epidemiology. *Methods in social epidemiology*, *1*, 370–393.

Park, K., Ewing, R., Scheer, B. C., & Khan, S. S. A. (2018). Travel Behavior in TODs vs. Non-TODs: Using Cluster Analysis and Propensity Score Matching. *Transportation Research Record*, *2672*(6), 31–39. https://doi.org/10.1177/0361198118774159

Patnala, P. K., Parida, M., & Chalumuri, R. S. (2023). Gender differentials in travel behavior among TOD neighborhoods: Contributions of built environment and residential self-selection. *Travel behaviour and society/Travel behaviour society*, *31*, 333–348. https://doi.org/10.1016/j.tbs.2023.01.005

Poorthuis, A., & Zook, M. (2023). Moving the 15-minute city beyond the urban core: The role of accessibility and public transport in the Netherlands. *Journal of Transport Geography*, *110*, 103629. https://doi.org/10.1016/j.jtrangeo.2023.103629

Pot, F. J., Koster, S., & Tillema, T. (2023). Perceived accessibility and residential self-selection in the Netherlands. *Journal of transport geography*, *108*, 103555. https://doi.org/10.1016/j.jtrangeo.2023.103555

Profillidis, V., & Botzoris, G. (2018, October). *Modeling of transport demand*. Elsevier.

ProRail. (n.d.-a). Innoveren. Retrieved January 18, 2024, from https://www.prorail.nl/toekomst/innoveren

ProRail. (n.d.-b). Mobiliteit van morgen. Retrieved January 18, 2024, from https://www.prorail.nl/over-ons/wat-doet-prorail/mobiliteit-van-morgen

Puylaert, S., Clerx, W., & Veurink, A. (2022, October). *Binnensteden zijn anders; dataonderzoek naar de mobiliteitstransitie in de stad* (tech. rep.). Colloquium Vervoersplanologisch Speurwerk. https://www.cvs-congres.nl/e2/site/cvs/custom/site/upload/file/paper_search/2022/cvs_162_binnensteden_zijn_anders_dataonderzoek_naar_de_mobiliteitstransitie_in_de_stad_1_2022.pdf

Rijkswaterstaat. (2022a). *Nationaal Wegenbestand - Parkeervlakken 2022-9-1*. https://downloads.rijkswaterstaatdata.nl/wkd/Parkeervlakken/01-09-2022/

Rijkswaterstaat. (2022b). *Nationaal Wegenbestand - Wegbreedte 2022-9-1*. https://downloads.rijkswaterstaatdata.nl/wkd/Wegbreedte/01-09-2022/

Rijkswaterstaat. (2022c). *Nationaal Wegenbestand - Wegcategorisering 2022-7-1*. https://downloads.rijkswaterstaatdata.nl/wkd/Wegcategorisering/01-07-2022/

Rijkswaterstaat. (2022d). *Nationaal Wegenbestand - Wegvakken 2018-10-1*. https://downloads.rijkswaterstaatdata.nl/nwb-wegen/geogegevens/shapefile/Nederland_totaal/01-10-2018/Wegvakken/

Rijkswaterstaat. (2022e). *Nationaal Wegenbestand - Wegvakken 2022-9-1*. https://downloads.rijkswaterstaatdata.nl/nwb-wegen/geogegevens/shapefile/Nederland_totaal/01-09-2022/Wegvakken/

Rijkswaterstaat, V. e. L., Water, & Smit, R. (2018a, June). *Memo Gebruik van OViN voor analyses op werkdagen* (tech. rep.).

Rijkswaterstaat, Water, Verkeer en Leefomgeving. (2017). *LMS Zonenummering 2018 met PC4*.

Rijkswaterstaat, Water, Verkeer en Leefomgeving. (2018b). *LMS Stationskenmerken*.

Rijkswaterstaat, Water, Verkeer en Leefomgeving. (2018c). *SES HB-matrices LMS - all modes*.

Rijkswaterstaat, Water, Verkeer en Leefomgeving. (2020). *Sociaaleconomische gegevens en shapefile LMS 2018*.

Rijkswaterstaat, Water, Verkeer en Leefomgeving. (2021a, May 26). *Documentatie van gm4 deel d10: Begrippen en definities*.

Rijkswaterstaat, Water, Verkeer en Leefomgeving. (2021b, May 26). *Documentatie van gm4 deel d2: Systeemstructuur*.

Rijkswaterstaat, Water, Verkeer en Leefomgeving. (2021c, May 26). *Documentatie van gm4 deel d4-1: Programma quad*.

Rijkswaterstaat, Water, Verkeer en Leefomgeving. (2021d, May 26). *Documentatie van gm4 deel d4-2: Programma carmod*.

Rijkswaterstaat, Water, Verkeer en Leefomgeving. (2021e, May 26). *Documentatie van gm4 deel d5: De bereikbaarheidsmodule*.

Rijkswaterstaat, Water, Verkeer en Leefomgeving. (2021f, May 26). *Documentatie van gm4 deel d7: De groeifactormodule*.

Rijkswaterstaat, Water, Verkeer en Leefomgeving. (2021g, May 26). *Documentatie van gm4 deel d7-1: Programma ses*.

Rijkswaterstaat, Water, Verkeer en Leefomgeving. (2021h, May 26). *Documentatie van gm4 deel d7-2: Programma secdest*.

Rijkswaterstaat, Water, Verkeer en Leefomgeving. (2021i, May 26). *Documentatie van gm4 deel d7-3: Programma nhbtrips*.

Rijkswaterstaat, Water, Verkeer en Leefomgeving. (2021j, May 26). *Documentatie van gm4 deel d7-5: Programma pivot*.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rubin, O., Mulder, C. H., & Bertolini, L. (2014). The determinants of mode choice for family visits – evidence from Dutch panel data. *Journal of Transport Geography*, *38*, 137–147. https://doi.org/10.1016/j.jtrangeo.2014.06.004

Schwanen, T., Dijst, M., & Dieleman, F. M. (2002). A Microlevel Analysis of Residential Context and Travel Time. *Environment and Planning A: Economy and Space*, *34*(8), 1487–1507. https://doi.org/10.1068/a34159

Sikder, S., Pinjari, A. R., Srinivasan, S., & Nowrouzian, R. (2013). Spatial transferability of travel forecasting models: A review and synthesis. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, *5*(2-3), 104–128. https://doi.org/10.1007/s12572-013-0090-6

Snelder, M., & Vonk Noordegraaf, D. (2022, February). *Review LMS/NRM* (tech. rep. No. TNO 2022 R10341). TNO. https://publications.tno.nl/publication/34639925/oQkGSx/TNO-2022-R10341.pdf

Stead, D., & Marshall, S. (2001). The Relationships between Urban Form and Travel Patterns. An International Review and Evaluation. *European Journal of Transport and Infrastructure Research*. https://doi.org/10.18757/ejtir.2001.1.2.3497

Sung, H., & Eom, S. (2024). Evaluating transit-oriented new town development: Insights from Seoul and Tokyo. *Habitat international*, *144*, 102996. https://doi.org/10.1016/j.habitatint.2023.102996

Susilo, Y. O., & Maat, K. (2007). The influence of built environment to the trends in commuting journeys in the Netherlands. *Transportation*, *34*(5), 589–609. https://doi.org/10.1007/s11116-007-9129-5

Taboga, M. (2021). "Dummy variable", Lectures on probability theory and mathematical statistics. https://www.statlect.com/fundamentals-of-statistics/dummy-variable

Taubenböck, H., Droin, A., Standfuß, I., Dosch, F., Sander, N., Milbert, A., Eichfuss, S., & Wurm, M. (2022). To be, or not to be 'urban'? a multi-modal method for the differentiated measurement of the degree of urbanization. *Computers, Environment and Urban Systems*, *95*, 101830. https://doi.org/https://doi.org/10.1016/j.compenvurbsys.2022.101830

Thao, V. T., & Ohnmacht, T. (2020). The impact of the built environment on travel behavior: The Swiss experience based on two National Travel Surveys. *Research in Transportation Business Management*, *36*, 100386. https://doi.org/10.1016/j.rtbm.2019.100386

Thomas, T., & Tutert, S. (2013). An empirical model for trip distribution of commuters in the netherlands: Transferability in time and space reconsidered. *Journal of Transport Geography*, *26*, 158–165. https://doi.org/10.1016/j.jtrangeo.2012.09.005

Train, K. (2009). *Discrete Choice Methods with Simulation* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511805271

University of Groningen Geodienst. (2021). *Openbaar vervoer Nederland - Haltes*. https://hub.arcgis.com/datasets/RUG::openbaar-vervoer-nederland/about?layer=0

Van Acker, V., Mokhtarian, P. L., & Witlox, F. (2014). Car availability explained by the structural relationships between lifestyles, residential location, and underlying residential and travel attitudes. *Transport policy*, *35*, 88–99. https://doi.org/10.1016/j.tranpol.2014.05.006

Van Acker, V., Mokhtarian, P. L., & Witlox, F. (2011). Going soft: on how Subjective Variables Explain Modal Choices for Leisure Travel. *European Journal of Transport and Infrastructure Research*. https://doi.org/10.18757/ejtir.2011.11.2.2919

Van Acker, V., Van Wee, B., & Witlox, F. (2010). When Transport Geography Meets Social Psychology: Toward a Conceptual Model of Travel Behaviour. *Transport Reviews*, *30*(2), 219–240. https://doi.org/10.1080/01441640902943453

Van De Coevering, P., & Schwanen, T. (2006). Re-evaluating the impact of urban form on travel patternsin Europe and North-America. *Transport Policy*, *13*(3), 229–239. https://doi.org/10.1016/j.tranpol.2005.10.001

Van Der Hoorn, A. (1979). Travel behaviour and the total activity pattern. *Transportation*, *8*(4), 309–328. https://doi.org/10.1007/bf00167986

Van Leeuwen, J., & Venema, N. (2023, May). Statistische gegevens per vierkant en postcode 2022-2021-2020-2019. https://www.cbs.nl/nl-nl/longread/diversen/2023/statistische-gegevens-per-vierkant-en-postcode-2022-2021-2020-2019

Van Nes, R., & De Jong, G. (2020, January). Transport models. In *Advances in transport policy and planning* (pp. 101–128). https://doi.org/10.1016/bs.atpp.2020.08.001

Vos, J. D. (2015). The influence of land use and mobility policy on travel behavior: A comparative case study of flanders and the netherlands. *Journal of Transport and Land Use*, *8*(1), 171–190. Retrieved June 11, 2024, from http://www.jstor.org/stable/26202708

Yu, A. C. H., & Higgins, C. D. (2024). Travel behaviour and the 15-min City: Access intensity, sufficiency, and non-work car use in Toronto. *Travel behaviour and society/Travel behaviour society*, *36*, 100786. https://doi.org/10.1016/j.tbs.2024.100786

A

Preliminary literature review: Research gap

This appendix presents the preliminary literature review that was done before the start of this thesis to find the research gap. A summary of the most important findings can be found in section 1.3.

A.1. Methodology

This subsection will describe the methodology that was used to find out what research has been done on region specific factors in transport modelling and to find the research gap. For this, several search engines have been used (ScienceDirect, Scopus and Google Scholar). Different (combinations of) keywords were inserted in those search engines to search for an initial set of papers. This initial search consisted of the following words:

> *("traffic forecasting" OR "traffic modelling") AND ("degree of urbanisation" OR "degree of urbanization")*

These keywords gave 49 hits in Google Scholar, 19 hits in ScienceDirect and 0 hits in Scopus. This could be explained by the fact that Google Scholar and ScienceDirect search the whole documents for keywords, while Scopus only searches through the abstracts, article titles and keywords. By removing "degree of" from the keywords while searching in Scopus, it gave 19 hits. These results were evaluated based on the title and abstract and less than 5 articles seemed partly relevant. There was one article that stood out. It used the keyword: *"spatial transferability"*. In the context of transport modelling it means the ability of a model that was trained for one region, to model transport in a different region (Sikder et al., 2013). This study focuses on the ability of one large model to differentiate between regions and not on the ability of one smaller model to be applied to a different region, so it is not completely relevant. However, it seems like a good starting topic of this literature review, because it will provide more information about the effects different regions have on transport models.

This gave more new and more specific keywords to try out in the above mentioned search engines, e.g. *"spatial transferability" AND ("transport" OR "traffic") AND "discrete choice" AND ("modelling" OR "forecasting")* or small variations on these keywords. This gave a new set of potential papers to study. Another source of papers that was later used is the CVS database. This database contains many semi-scientific Dutch papers related to transportation. All these papers have both many useful references themselves and are being cited to by other papers, which also provided important sources.

A.2. Spatial transferability

To the author's knowledge, there are no studies that specifically focus on the ability of one large transport model to capture the differences in travel behaviour in smaller sub regions with a similar or different degree of urbanisation.

However, there are many studies that focus on spatial transferability, as described in the methodology of this section. These studies often focus on a certain city or a region and compare whether a

transport forecasting model trained for that specific region can also be applied to another region with similar characteristics. Even though this topic is not directly applicable to the topic of this thesis, it could potentially provide useful information and tools to get more knowledge about the difference in travel behaviour in different regions. A lack of spatial transferability of a certain region and a certain model, implies that there are region specific characteristics that the model was not able to capture.

McArthur et al. (2011) fits a gravity model to estimate commuter trips in three different regions in Norway. All three regions are similar in socio-economic factors and in the quality of the available public services. After that, the model that was fitted to one region, was applied to the other two regions and the results were analysed. This study showed that in some cases applying a model to a different region gave good results, while in other cases it did not. However, in each case the predictions of the other models were in the same order of magnitude as the original model, which could be good enough in some cases. The problem that this paper mentions is that it is not possible in advance to know if a certain model is transferable to another region or not. The characteristics that separated the similar looking regions from each other were unknown.

Thomas and Tutert (2013) presents similar research on fitting a gravity model to commuter trips, based on Dutch Travel surveys. They also found that the model is not spatially transferable and the travel behaviour in different city sizes do not change the way it would be expected based on the gravity model. This study admits that the factors included in their model are far from complete and urban and regional factors should be included for better results.

The two studies mentioned above only focused on commuter flows while using a gravity model. Sikder et al. (2013) conducted an extensive literature review on spatial transferability of travel demand models on a regional level (not nation wide). It focuses mostly on tour-based and ABMs. Even though the LMS is not necessarily an ABM, this method still provides clues on how to improve the LMS. The study suggests that there can be differences between contexts on a more theoretical level, that prevent transferability. For example, normally it is assumed that travellers want to maximize utility, but that might not be true in each context. Different decision-making theories might apply. However there does not yet exist empirical evidence on the effect of different decision-making theories on the transferability of models. It also suggests that the choice of a specific model (e.g. nested logit or probit) brings in additional assumptions. Often, the choice of a specific model structure is based on underlying theory and assumptions. These different assumptions can introduce errors, which might vary in different contexts. Different regions might require different model structures. This suggests that it might be possible that people in more urban regions make transport choices in a different way than people in more rural regions. That would mean that the current model structure might not be the optimal structure for each region (e.g. the nested logit model in the GM determines first the mode and then the destination. Perhaps this is a logical order of choices in rural regions, but not in urban regions, or the reverse). It would be interesting to research this topic more in the thesis.

Besides the transferability of the model structure, according to Sikder et al. (2013) there is some evidence that mode choice and location choice are the least transferable elements of a model, which suggests that these elements will need some extra attention to model in different regions. Linh et al. (2019) found that in the case of transferring a model that was made for Belgium to a region in Vietnam, that mode choice and location choice are difficult to transfer because of the differences in land-use, location preferences and mode availability between the two different regions.

Sikder et al. (2013) states that it would be useful to investigate if it is possible to identify several region categories to standardize ABM frameworks. They give some suggestions of factors to include in these categories: the size of the region, the socio-demographics, available modes, land-use, the features of the transportation network and policies with regard to transportation planning. Data about these regions should then be collected and combined to form the different categories. When doing this, it is still important to acknowledge that there are always regional differences, also within regions that are part of the same category. This can be done by adding additional explanatory variables.

Currently in the LMS, regions are mainly categorized based on their degree of urbanisation, often combined with other factors. Based on Sikder et al. (2013) however, it would be interesting to research if it is possible to define categories that are based on much more factors than only the degree of urbanisation.

A.3. Hypothesis and research gap

It is often unclear whether models are spatially transferable between different regions or not. This could be due to the fact that the characteristics that separate those regions are unknown (McArthur et al., 2011). Factors, with some evidence, that make spatially transferring a model more difficult are mode choice and location choice due to differences in land-use, location preferences and mode availability between regions (Sikder et al., 2013; Linh et al., 2019).

Two main results from this literature review will be interesting to look further into. Sikder et al. (2013) suggests that it is unclear if the same model structure can always be applied to different regions and recommends to investigate whether it is possible to identify different region categories that separate different kind of regions from each other.

After consulting with experts, the first idea would be theoretically and practically very difficult to research and implement. That is why the second idea will be further researched in this thesis. It will be interesting to research if it is possible to introduce a new variable that can replace the degree of urbanisation in the LMS and is better in distinguishing differences in travel behaviour between different regions. This variable should also be able to be used in other transport models besides the LMS, that currently mainly rely on variables related to the population density to distinguish between regions. The degree of urbanisation is currently used in many different places in the LMS. It is used as variables in the mode-part of day-destination discrete choice model, but it is also used on a higher level (e.g. when determining the population distribution for a zone). This will be further elaborated in section 3.3. The analysis done in this thesis will focus mainly on the mode choice and how the degree of urbanisation effects this. Advice given on possible new variables to replace the degree of urbanisation will, consequently, also focus primarily on mode choice and less on how the degree of urbanisation might affect the LMS on a more structural level.

To conclude, not much research exists about modelling travel behaviour between different regions in one large model. The lack of research in this area can be combined with the research gaps found in the literature about spatial transferability of transport models, which gives the research gap for this thesis.

B

Overview relevant coefficients from LMS documentation

This appendix gives an overview of the coefficients in the LMS that are related to the spatial environment. This is further elaborated in section 3.3. Tables B.1 and B.2 show the different motives and modes that are used in the LMS to determine the frequency and MDD choice. The numbers and abbreviations in those tables are used in the remaining of the appendix to indicate the different motives and modes.

Table B.3 shows the coefficients of the car ownership module; table B.4 shows the coefficients in the frequency model; and table B.5 shows the coefficients of the MDD choice model. Only variables that are related to the spatial environment are shown in this appendix. For the full list of coefficients, see RWS WVL (2021a).

Table B.1: Overview of the different motives that are used in the LMS. This table is based on table 4.1 and table 5.10 from RWS WVL (2021g).

Motive	Number
Home-education for students	1a
Home-education for fulltime workers, parttime workers and not-students	1b
Home-work for fulltime workers (30+ hours)	2a
Home-work for parttime workers, students and not-students	2b
Home-business	3
Home-shopping	4
Home-other	5
Work-business	6
Work-other	7
Child-education (<12 years)	8
Child-shopping (<12 years)	9
Child-other (<12 years)	10

Table B.2: Overview of the different modes that are used in the LMS. This table is based on table 5.10 from RWS WVL (2021g).

Mode	Abbreviation
Car driver	AB
Car passenger	AP
Train	TR
BTM	BTM
Tram/Metro	TM
Bus	B
E-bike	Eb
Bike	F
Walking	L

Table B.3: Car ownership coefficients that are related to the spatial environment form the CARMOD module in the LMS. Based on table 2.1 from RWS WVL (2021a).

Description	Alternatives: number of cars			
	0	1	2	3+
Disposable income after subtracting costs for 1 or more cars, if houshold has a DU of 1, 2 or 3		0.265	1.307	1.566
Disposable income after subtracting costs for 1 or more cars, if houshold has a DU of 4, 5 or 6		0.2179	1.203	1.418
Population density of all zones within 1 km of the centroid of a zone		-0.00283	-0.00588	-0.00888
Job density of all zones within 1 km of the centroid of a zone		-0.00191	-0.00487	-0.00468
Job density of all zones within 5 km of the centroid of a zone		-0.00735	-0.01803	-0.02554
Parking fee per hour in Euros	0.1653	0.06233		
Average maximum number of parking permits per household (if parking permits <4)		0.1197	0.2927	0.1309
Dummy variable of 1 for a zone without parking limitations (if parking permits >= 4)		0.7045	1.5	0.8529
Dummy variable of 1 if household is in a zone with a DU of 3 or 4	-0.1726			
Dummy variable of 1 if household is in a zone with a DU of 1, 2 or 3	-0.9327	-0.3614	-0.2167	
The ratio of agricultural jobs (no. of agricultural jobs / total no. of jobs)	-2.12			

Table B.4: Coefficients for the frequency model from the SES module in the LMS that are related to the spatial environment. This table is based on table 2.33 and 2.34 from RWS WVL (2021a). The 'model' column shows in which part of the model the coefficient is used. '0/1+' refers to the first step (determining if 0 or 1+ trips are made) and 'stop/repeat' refers to the second step (determining if an additional trip is made).

Description	Model	Coefficients for each motive									
		1a	1b	2a	2b	3	4	5	8	9	10
No car, no driver's licence	0/1+				0.8987			0.09483			
Driver's licence, but no car	0/1+	-0.5575					-0.2199				
Car, but no driver's licence	0/1+	-0.1761			0.8489		0.06989				
	stop/repeat	0.08759						0.341			
Car under competition	0/1+	-0.1778				-0.4115		-0.1453			
	stop/repeat			-0.2899				0.1992			
Car freely available	0/1+			-0.03371		-0.2173		-0.07931			
	stop/repeat							0.3939			
DU of 1 or 2	0/1+			0.09608			0.06893		-0.09873	-0.1274	-0.117
	stop/repeat			-0.8698	-0.4145			-0.2358	-1.368	-1.447	-0.139
DU of 3 or 4	0/1+	0.1109								-0.2739	-0.09491
	stop/repeat			-0.3393					-0.9644	-1.274	
DU of 1	0/1+					-0.2875					
DU of 1, 2 or 3	stop/repeat					-0.6346					
DU of 4 or 5	stop/repeat					-0.4868					
Logsum: measure of accessibility of a zone and the type of person	0/1+		-0.2256			-0.5597	-0.1144	-0.2193			
	stop/repeat				-0.1607		-0.1605	-0.7297			

Table B.5: Coefficients for the mode destination and part of day choice related to the spatial environment from the SES module in the LMS. This table is based on table 2.35 and 2.36 from RWS WVL (2021a).

Description	Mode	Coefficients for each motive									
Travel time		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Parking fee	AB	-0.00393	-0.00197	-9.95E-04	-0.00432	-9.57E-04	-5.49E-04				
Travel time car	AB, AP	-0.7566	-0.05852	-0.04581	-0.06507	-0.03116	-0.0629	-0.04331	-0.03903	-0.08761	-0.08761
Travel time as car passenger	AP	-0.03389	-0.02309	-0.07541	-0.09612	-0.04597		-1.20E-05	-0.05338		
Travel time access/egress BTM	TM, B	-0.07422	-0.06551			-0.05033					
In-vehicle time tram/metro	TM	-0.037	-0.01701			-0.01541					
In-vehicle time bus as access/egress tram/metro	TM	-0.05709	-0.05626			-0.0416					
Generalized travel time bus	B	-0.03383	-0.03278			-0.02546					
Generalized travel time BTM	TM, B			-0.07788	-0.05109			-0.04409	-0.03472	-0.04585	-0.04585
Travel time bike	F	-0.1337	-0.1159		-0.2165	-0.1347				-0.1098	-0.1098
Travel time E-bike	Eb	-0.1279	-0.09084		-0.1771	-0.1023					
Travel time E-bike/bike	F, Eb			-0.4015			-0.2544	-0.1345	-0.1352		
Travel time walking	L	-0.1123	-0.08199	-0.1542	-0.1219	-0.07203		-0.02912	-0.08457	-0.04715	-0.04715
Person		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Car under competition	AB	-0.7799	-1.297	-0.8539	-0.6153	-0.6137					
Driver's licence, but no car	AB	-3.816	-3.737	-2.111	-3.271	-2.758					
Car freely available	AP				-1.168	-0.06523					
Car under competition	AP				-0.6332						
Car, but no driver's licence	AP					0.4531					
No car	AP	-1.076	-0.5752		-1.716	-0.3326			-1.786	-1.396	-1.396
No car	TR									1.734	1.734
No car and/or no driver's licence	TM, B		0.3563								
No car	TR, TM, B		0.4361		0.9154	1.323			1.396		
Car under competition	TR, TM, B					-0.3475					
Size		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Total no. of jobs, including self-employed	All						1.608				
No. of students in special education	All								-0.8358		
No. of jobs in service sector	All				-4.444	-6.73E-04					
No. of jobs in retail sector	All			1.129		1.873				2.145	2.145
No. of jobs in agriculture	All			0.9595							
Population density destination zone	All					-0.00634			-0.00432	-0.00727	-0.00727
Population density origin zone	All					-0.00188					
Zonal		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Distance coefficient if no worker	All		-0.01445	-0.01273							
Distance coefficient if part-time worker	All		-0.01403								
Distance coefficient if full-time worker	All	0.01183									
Distance coefficient if primary or lower education	All		-0.00615								
Distance coefficient if higher education	All		0.00581								
Distance coefficient if age <18	All	-0.00738									
Distance coefficient if age >54	All	-0.00351									
Job density >75 jobs/ha	All		0.1447	-0.4006		0.2732					
Job density >75 jobs/ha	AB		-0.4314								
DU origin zone 4	AB		-0.1281								
DU origin zone 5 or 6	AB		-0.3533			-0.5144					
DU origin zone 4, 5 or 6	AB				-0.2211						
DU origin zone 5 or 6	AP					-0.3514					
DU origin zone 5 or 6	TR		0.3886								
DU destination zone 4	AB		-0.2286		-0.1894	-0.2776					
DU destination zone 5 or 6	AB		-0.3122		-0.2725	-0.4339					
DU destination zone 4, 5 or 6	AB	-0.3459						-0.5907			
DU destination zone 4	AP	-0.8159	-0.3138		-0.2464	-0.1325					
DU destination zone 5 or 6	AP	-1.547	-0.732		-0.8869	-0.2004					
DU destination zone 4, 5 or 6	AP									-0.2122	-0.2122
DU destination zone 5 or 6	TR			0.7921		0.5886					
DU destination zone 5 or 6	F, Eb	-0.8733	-0.1016		-0.199	-0.262					
DU destination zone 5 or 6	L	1.016	0.556		-0.8959	0.4337		16.18		0.604	0.604
Euclidean distance >80 km	All					3.364					
Share of high education jobs	All		2.957	2.265							
Share of high education jobs if high educated	All		1.72								
Share of high education jobs if medium educated	All		-0.3056								
Share of high education jobs if low educated	All		-1.398								
Train		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Car driver access constant	AB	-1.995	-2.654	-2.756	-2.963	-3.345	-2.711				
Car passenger access constant	AP	-2.621	-3.683	-3.658	-2.253	-2.544	-3.256			3.616	3.616
Tram/metro access constant	TM	-0.569	-1.606	-1.879	-1.614	-1.595	-1.813			3.796	3.796
Tram/metro egress constant	TM	-1.968	-1.705	-2.568	-1.398	-1.473	-2.751			0.3781	0.3781
ASC for shared bike	F	-6.868	-4.346	-5.048	-5.64	-4.209	-99			-99	-99
Bus acess constant	B	-0.7933	-1.869	-2.006	-1.803	-1.831	-1.838			3.893	3.893
Bike access constant	F	-0.6364	-1.147	-1.457	-2.06	-2.157	-1.376			4.087	4.087
Car passenger egress constant	AP	-6.19	-5.722	-4.698	-3.935	-2.308	-4.506			-1.026	-1.026
Bus egress constant	B	-1.986	-2.282	-2.809	-1.677	-1.756	-2.778			-0.74	-0.74
Bike egress constant	F	-4.418	-3.556	-7.297	-5.1	-4.614	-5.927			-3.342	-3.342
Student using BTM as access mode	BTM		0.7063			0.4952					
Student using bike as acces mode	F					0.376					
Worker using bike as access mode	F					-0.05851					
Worker using bike as egress mode	F					0.213					
Ratio for distance egress / total euclidean distance	BTM, TM, B	-2.309	-5.202	-1.414	-11.24	-6.239	-0.4257			0.2218	0.2218
Station choice		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Travel time car driver from origin to station	TR	-0.1998	-0.1708	-0.1641	-0.1447	-0.1743					
Travel time car passenger from origin to station	TR	-0.195	-0.2132	-0.1845	-0.173	-0.187			-0.1794	-0.1794	-0.1794
Station with DU of 6, and car passenger as access mode	TR			-2.536		-0.9279					
Parking fee origin, and car driver as access mode	TR	-0.00866	-0.01106		-0.00851	-0.01271					
Parking fee origin, and car passenger as access mode	TR	-0.00487	-0.00446	-0.00295	-0.00578	-0.00609			-0.00966	-0.00966	-0.00966
Cycle time from origin to station	TR	-0.1567	-0.1545	-0.1378	-0.1308	-0.1552			-0.1179	-0.1179	-0.1179
Walking time from origin to station	TR	-0.09972	-0.1053	-0.0862	-0.08049	-0.1074			-0.1036	-0.1036	-0.1036
Travel time car from station to destination	TR	-0.1681	-0.1904	-0.1605	-0.1499	-0.2055			-0.3138	-0.3138	-0.3138
Cycle time from station to destination	TR	-0.2114	-0.1252	-0.1094	-0.03017	-0.1275			-0.07546	-0.07546	-0.07546
Walking time from station to destination	TR	-0.1382	-0.09293	-0.08277	-0.09517	-0.1011			-0.105	-0.105	-0.105
Transfer penalty	TR	-0.04164	-0.04322	-0.03752	-0.03444	-0.04239			-0.03632	-0.03632	-0.03632
Service penalty	TR	-0.04613	-0.05011	-0.05263	-0.0204	-0.05168			-0.02766	-0.02766	-0.02766
In-vehicle travel time non IC-train	TR	-0.05877	-0.06817	-0.0479	-0.04767	-0.4185			-0.04847	-0.04847	-0.04847
In-vehicle travel time IC-train	TR	-0.05564	-0.06213	-0.03491	-0.03794	-0.03129			-0.04069	-0.04069	-0.04069
Car parking places/ no. of departing trains access station	TR	0.3076	0.4842	0.5239	0.2418	0.399					
Bike parking places/ no. of departing trains access station	TR	0.5372	0.5067	0.3788	0.6731	0.7406			2.768	2.769	2.770
Bike parking places/ no. of departing trains egress station	TR	1.159	0.6529	1.305		0.5922					
No. of retail jobs per km2 - access station	TR	0.00641	0.01165	0.01297	0.01295	0.01716			0.01488	0.01488	0.01488
No. of retail jobs per km2 - egress station	TR	0.01741	0.00602	0.01523	0.0034	0.02309			0.01628	0.01628	0.01628
Tram/metro access constant	TR	-2.234	-0.908			-2.92					
Tram/metro egress constant	TR		-2.683			-5.727					
In-vehicle time bus access/egress	TR	-0.0527	-0.06553	-0.05526	-0.04459	-0.05639			-0.0578	-0.0578	-0.0578
In-vehicle time tram/metro access/egress	TR	-0.05216	-0.06446	-0.04421	-0.03567	-0.04155			-0.04624	-0.04624	-0.04624
Walking time to BTM access/egress	TR	-0.09322	-0.09568	-0.07184	-0.05797	-0.08918			-0.07514	-0.07514	-0.07514
Initial waiting time BTM	TR	-0.07905	-0.0983	-0.08289	-0.06689	-0.08459			-0.0867	-0.0867	-0.0867
Transfer time BTM access/egress	TR	-0.07905	-0.0983	-0.08289	-0.06689	-0.08459			-0.0867	-0.0867	-0.0867
Number of transfers BTM access/egress	TR	-0.20026	-0.24901	-0.20999	-0.16944	-0.21428			-0.21964	-0.21964	-0.21964

C

Overview of nests in the LMS mode-destination-part of day choice

This appendix shows an overview of the different nests that are used in the MDD choice model of the LMS. See figures C.1 and C.2.

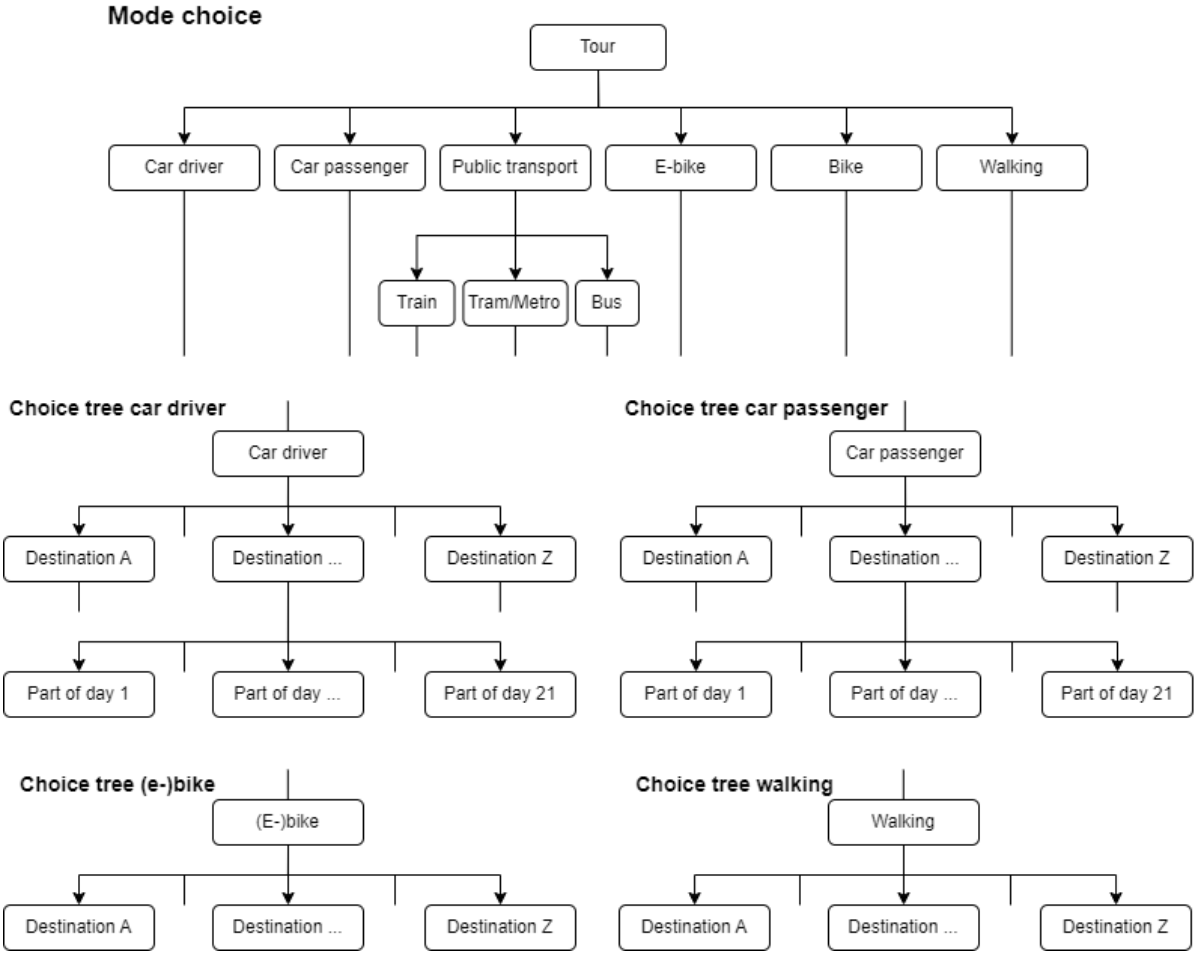

Figure C.1: Overview of all nests, except the public transport nests, that are used in the nested logit model of the MDD choice in the LMS. This figure is based on the figures 5.1-5.6 from RWS WVL (2021g).

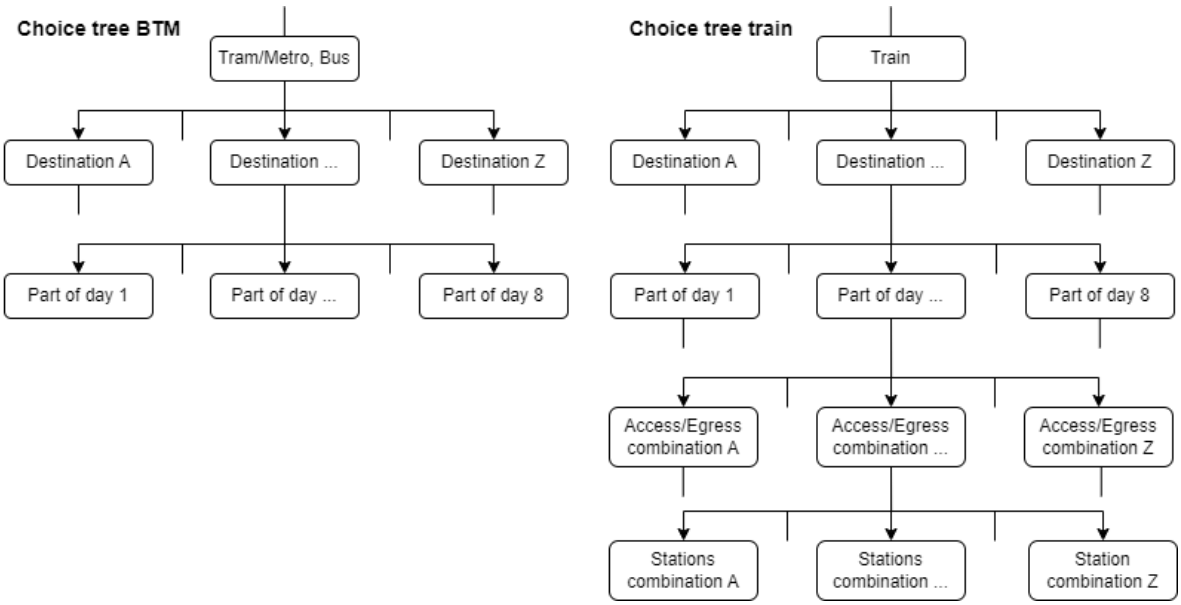

Figure C.2: Overview of the public transport nests that are used in the nested logit model of the MDD choice in the LMS. This figure is based on the figures 5.7 and 5.8 from RWS WVL (2021g).

D

Matching process of LMS zones with PC4, neighbourhoods and roads

This appendix will give an overview of of the process for matching the PC4 and neighbourhood zones with the LMS zones. For a summary of this process, see section 2.5.1 and 4.1.1.

In OViN the origin and destination locations of a trip are given on PC4 level. The LMS however uses its own zones, which are based on PC4 level, but are not exactly the same. To make the OViN trips comparable to the LMS data, each PC4 needs to matched to a LMS zone. Figure D.1 gives a comparison of the LMS zones and the PC4 zones. Especially in the northern provinces, the LMS zones are a lot larger than the PC4 zones. This figure shows that in most cases, several PC4 zones will belong to one LMS zone.

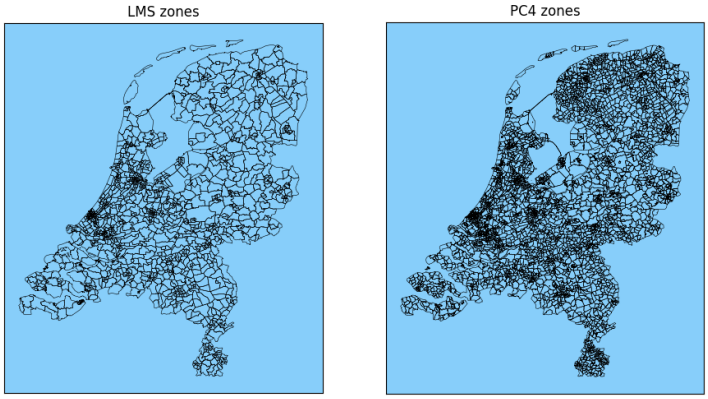

Figure D.1: Overview of zone borders for LMS (RWS WVL, 2020) and PC4 (CBS & ESRI Nederland, 2019) in the Netherlands. This figure shows that the LMS zones are often larger than PC4.

Rijkswaterstaat (RWS) provided a data file including documentation to match each PC4 zone with an LMS zone (RWS WVL, 2017). However in a few cases one PC4 zone was matched to two LMS zones or the PC4 zone was not included in the data file. A possible cause of this could be that this postal code did not appear in OViN 2015-2017 which were used in the LMS, but only appeared in the years 2013-2014. In total, 37 PC4s did not appear in this data file and 50 PC4s were matched to two different LMS zones.

In the case that a PC4 zone belonged to multiple LMS zones (see for example figure D.2), one of the LMS zones was chosen. A trip could not belong to multiple LMS zones at the same time, because some trips would be counted double in the statistics.

To match the remaining PC4 zones a small algorithm was written that matched the geographic centre of the PC4 zone with the nearest LMS zone. The shapefile with data for the PC4 zones is from 2019 (CBS & ESRI Nederland, 2019). The 37 PC4s without any match, were matched using a simple algorithm: the coordinates of the centroid of each PC4 zones were matched with the coordinates of the

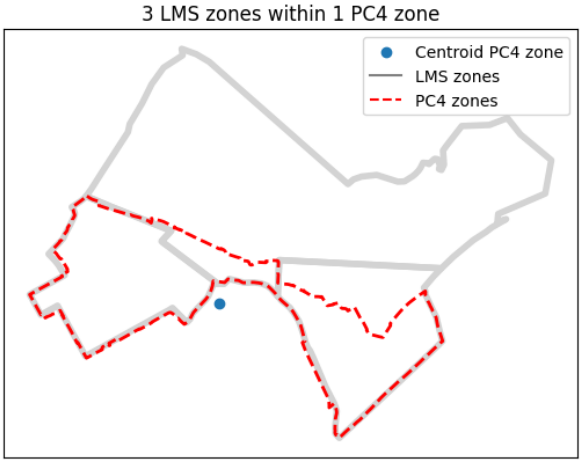

Figure D.2: Example of a PC4 zone that consists of more than 1 LMS zone. This map is based on RWS WVL (2020) and CBS & ESRI Nederland (2019).

LMS zone (i.e. the PC4 zone was matched with the LMS zone the centroid belonged to). This gave a good match for all the missing PC4s, except for 4. These postal codes did appear in the OViN dataset, but not in the PC4 dataset from 2019 (CBS & ESRI Nederland, 2019). These PC4s were presumably abolished in earlier or a mistake was made in the OViN data. Older PC4 datasets were searched. 2 of the PC4s existed around 2012. No record of the other PC4s was found. Trips with one of these 4 PC4s were removed from the OViN dataset. Finally, there were 152 OViN trips that missed the origin or destination zone. These trips were also removed from the dataset.

Data for land use existed only on neighbourhood level from 2017 and not on PC4 level (CBS & Kadaster, 2019). Neighbourhood zones are even smaller than PC4 zones on average. Often, many neighbourhood zones would fit in one LMS zone, see figure D.3. Again, several neighbourhood zones

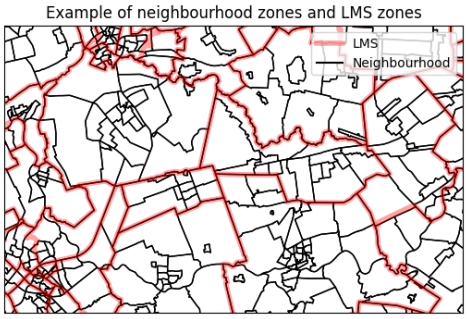

Figure D.3: Example of neighbourhoods zones in relation with LMS zones. This map is based on CBS & Kadaster (2019) and RWS WVL (2020).

were matched with multiple LMS zones, see figure D.4 for an example. In a 3 cases it was obvious to which LMS zone the neighbourhood belonged. The remaining 17 neighbourhood zones were assumed to belong fully to both LMS zones in the calculations.

After the matching, there were 2 LMS zones that did not have any land use data. This is undesirable because it is not possible to cluster zones with missing data. For these 2 zones, the closest neighbourhoods were identified manually.

Finally, road segments had to be matched with the corresponding LMS zone for the Design variables. To calculate the Design variables, shapefiles from the national road database (Nationaal Wegenbestand [NWB]) are used (RWS, 2022d; RWS, 2022e; RWS, 2022b; RWS, 2022c; RWS, 2022a). The matching was done using a *GeoPandas* function, in a similar way as when the PC4 and neighbourhood zones were matched with an LMS zone. A few sample zones were checked to see if the roads

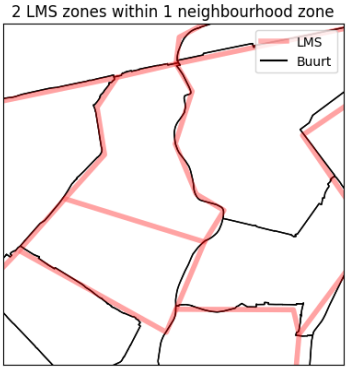

Figure D.4: Example of 2 LMS zones that fit into 1 neighbourhood zone. This map is based on CBS & Kadaster (2019) and RWS WVL (2020).

were matched with the right LMS zone. See for figure D.5 for an example of all road segments within one zone. Figure D.6 shows an example of all road segments in an LMS zone and their corresponding bike and pedestrian paths. This figure shows how the bicycle network is not yet completely added to the network. It mostly consists of separate road segments. Shared roads or painted bicycle lanes are not yet counted, which gives a distorted view of the bicycle network in the Netherlands.

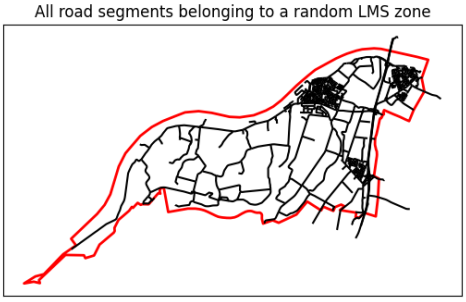

Figure D.5: Example of an LMS zone and its corresponding roads. The matching is fairly accurate, with only a few road segments crossing the border. This map is based on RWS WVL (2020) and RWS (2022d).

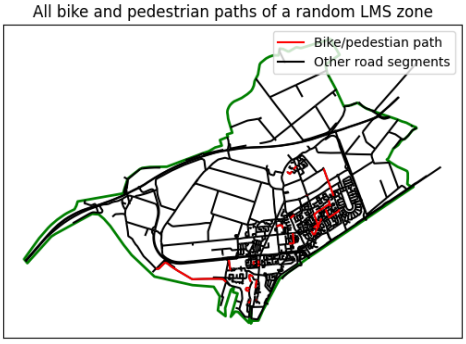

Figure D.6: Example of an LMS zone and its corresponding roads and bike roads. This map is based on RWS WVL (2020), RWS (2022e) and RWS (2022c).

E

Overview of all D-variables

This appendix will show the different maps that display the D-variables that were gathered. Maps that were already shown in the main text will not be shown again.

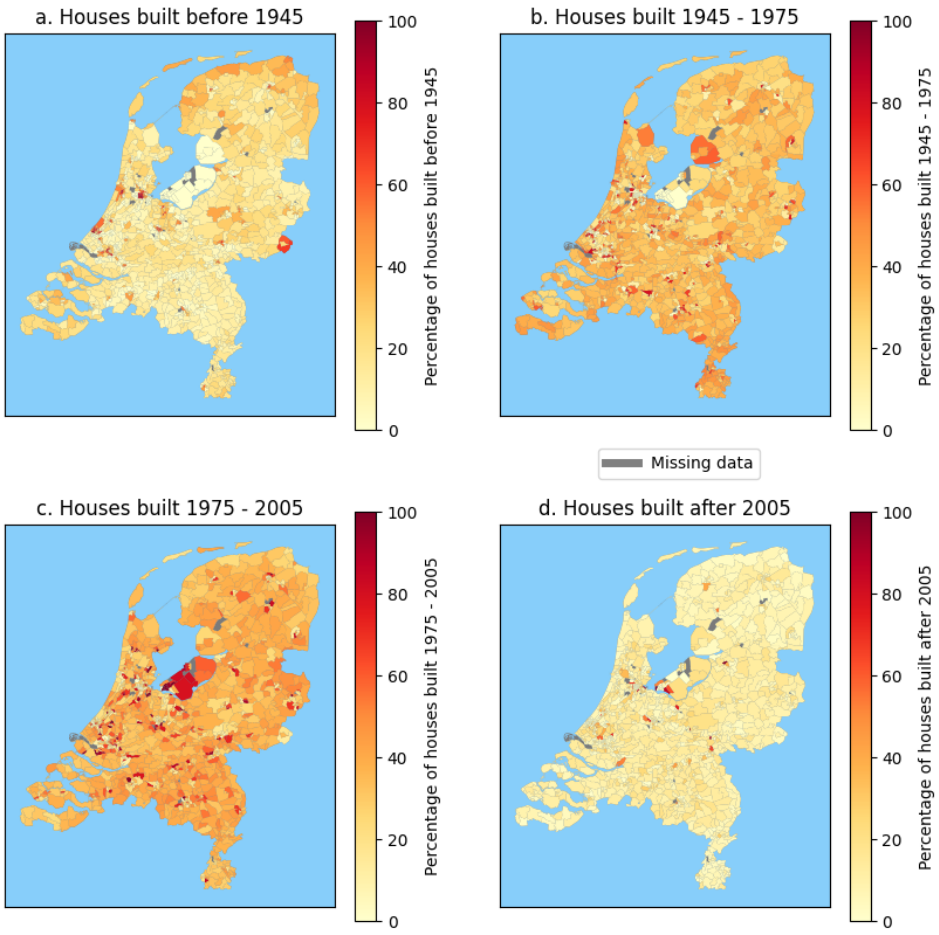

Figure E.1: Overview of the historical development in the Netherlands based on the built year of the houses. These variables are part of the Diversity variable. The shapes of the zones are based on RWS WVL (2020). The sources of the variables can be found in table 4.1.

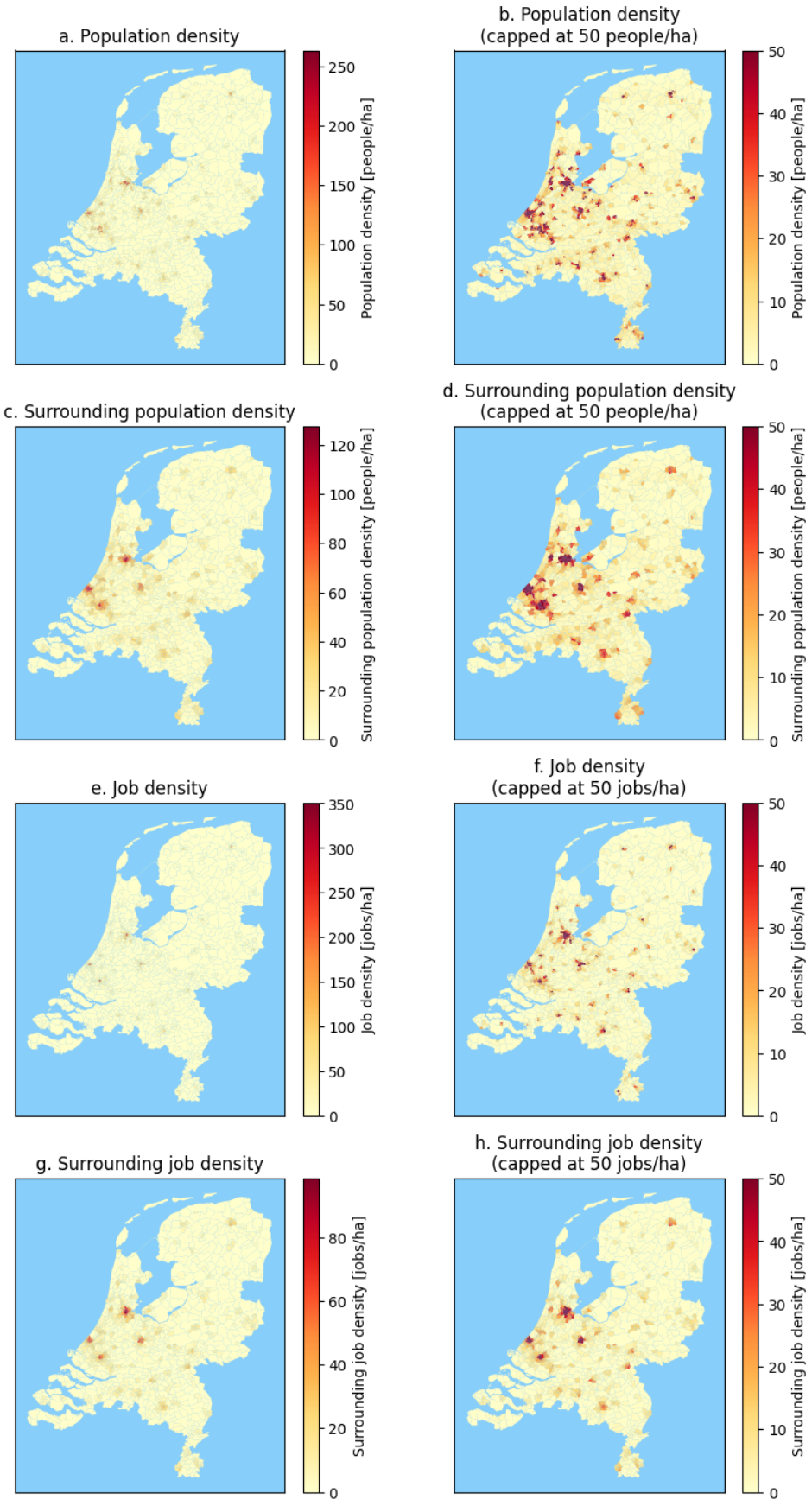

Figure E.2: Overview of all Density variables in the Netherlands, excluding the degree of urbanisation. Each variable is plotted a second time with a max of 50 people or jobs/ ha. This makes it easier to see patterns in less populated areas. The shapes of the zones are based on RWS WVL (2020). The sources of the variables can be found in table 4.1.

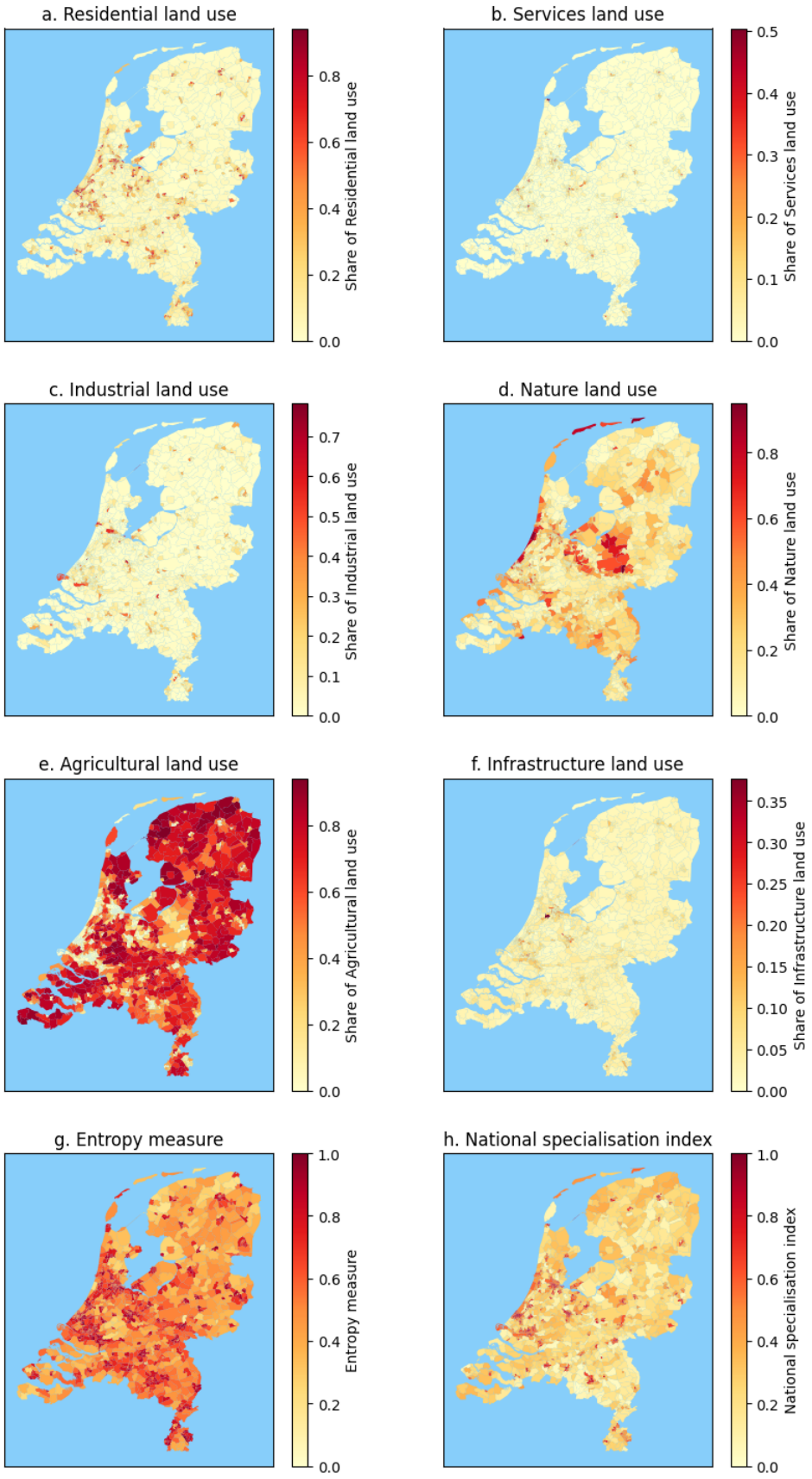

Figure E.3: Overview of the different land use types in the Netherlands. These variables are part of the Diversity variable. The shapes of the zones are based on RWS WVL (2020). The sources of the variables can be found in table 4.1.

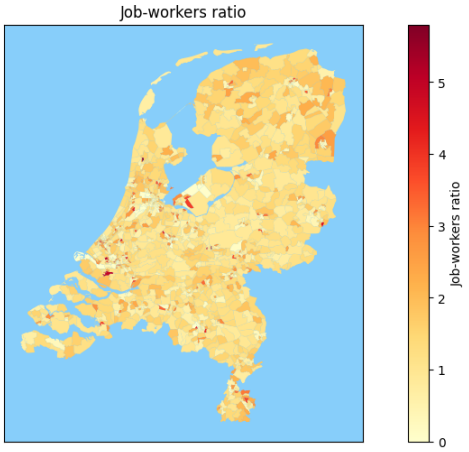

Figure E.4: Overview of the different jobs to workers ratio in the Netherlands. This variable are part of the Diversity variable. The shapes of the zones are based on RWS WVL (2020). The sources of the variables can be found in table 4.1.

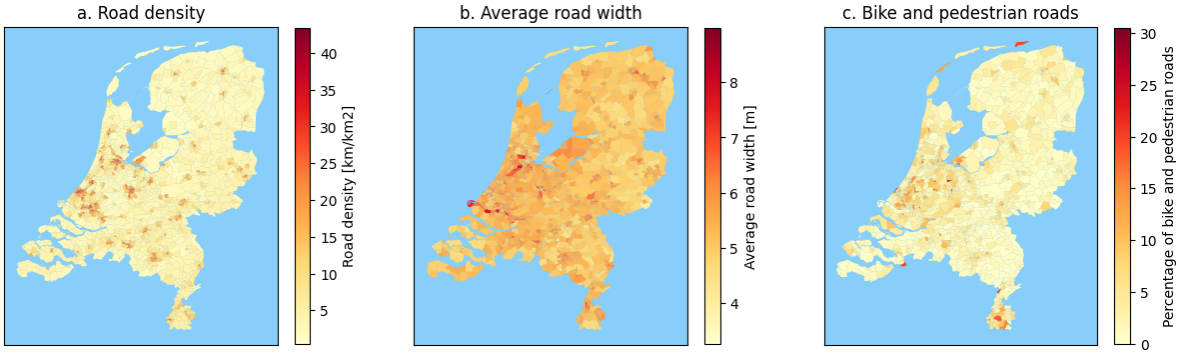

Figure E.5: Overview of the different Design variables in the Netherlands. The shapes of the zones are based on RWS WVL (2020). The sources of the variables can be found in table 4.1.

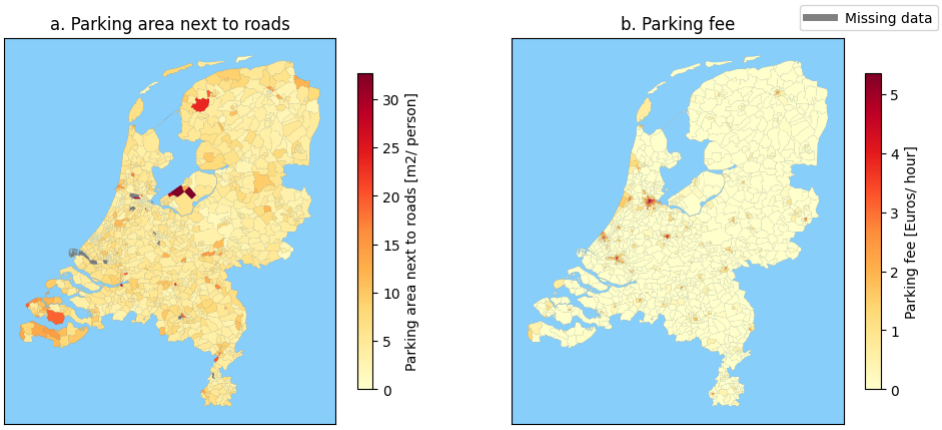

Figure E.6: Overview of the different Demand management variables in the Netherlands. The shapes of the zones are based on RWS WVL (2020). The sources of the variables can be found in table 4.1.

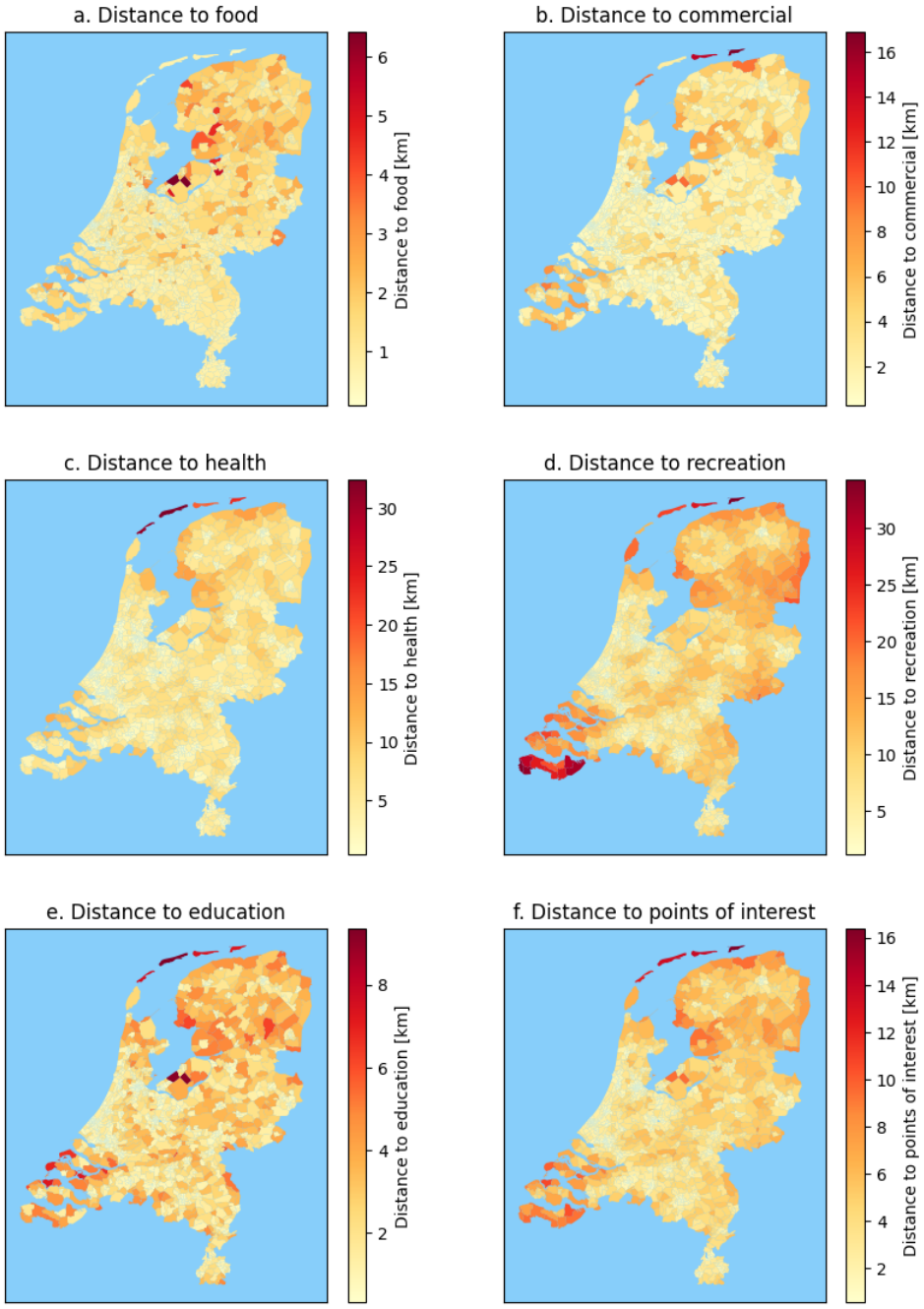

Figure E.7: Overview of the average distances to several points of interest. These variables are part of the Destination accessibility variable. The shapes of the zones are based on RWS WVL (2020). The sources of the variables can be found in table 4.1.

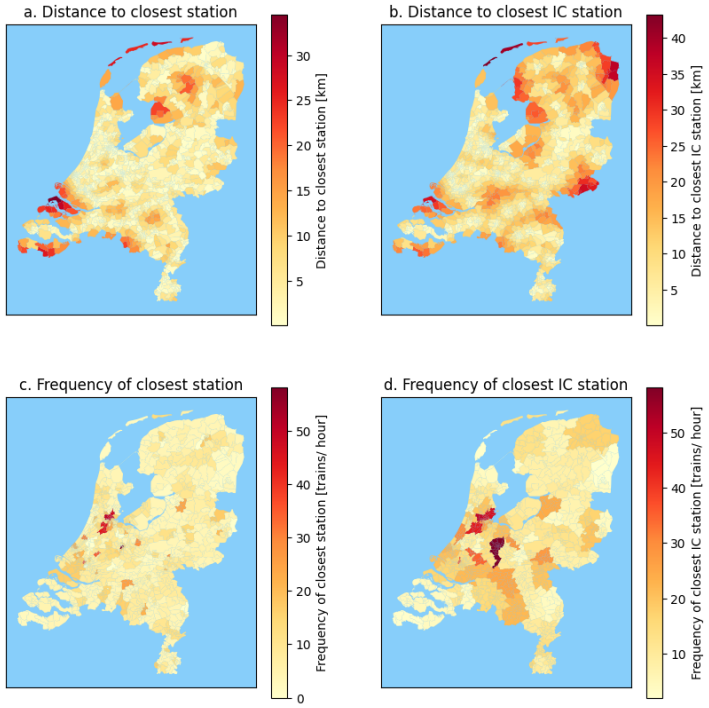

Figure E.8: Overview of the average distances and frequency to train stations. These variables are part of the Distance to transit variable. The shapes of the zones are based on RWS WVL (2020). The sources of the variables can be found in table 4.1.

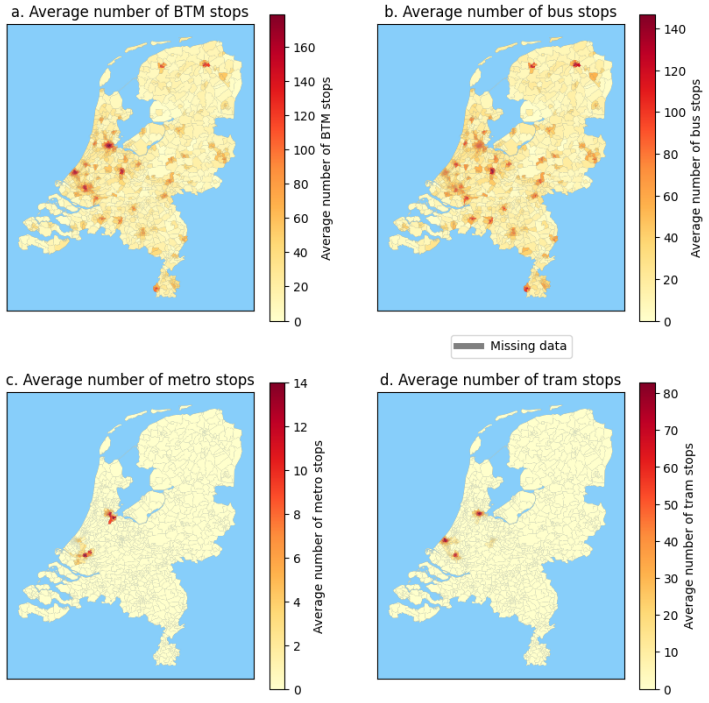

Figure E.9: Overview of the number of BTM stops within 2.5 km of the centroid each zone. These variables are part of the Distance to transit variable. The shapes of the zones are based on RWS WVL (2020). The sources of the variables can be found in table 4.1.

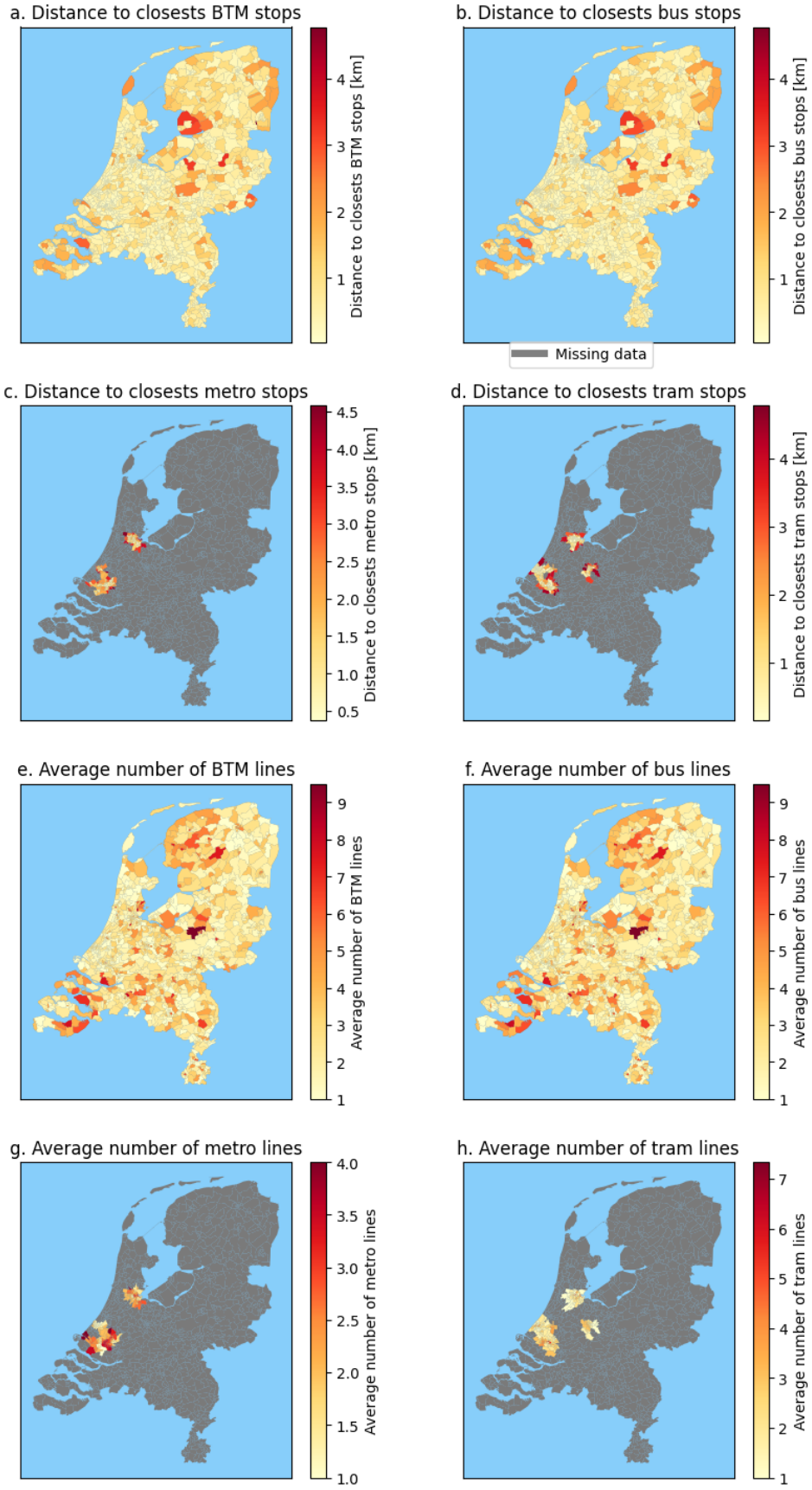

Figure E.10: Overview of variables related to the distance to BTM stops and the number of BTM lines. These variables are part of the Distance to transit variable. The shapes of the zones are based on RWS WVL (2020). The sources of the variables can be found in table 4.1.

F

Additional figures exploratory data analysis

This appendix shows some additional figures for the exploratory data analysis. All zones with less than 20 data points in OViN were filtered. Those zones have the same color as the background and can be recognized by a thin black line. In some cases the zone in OViN had more than 20 data points in total, but 0 trips with a certain mode. These zones were not filtered in the OViN and LMS plots, but resulted in missing values in the difference plots (right column). These zones in the difference plots are plotted in the same way as zones with less than 20 data points.

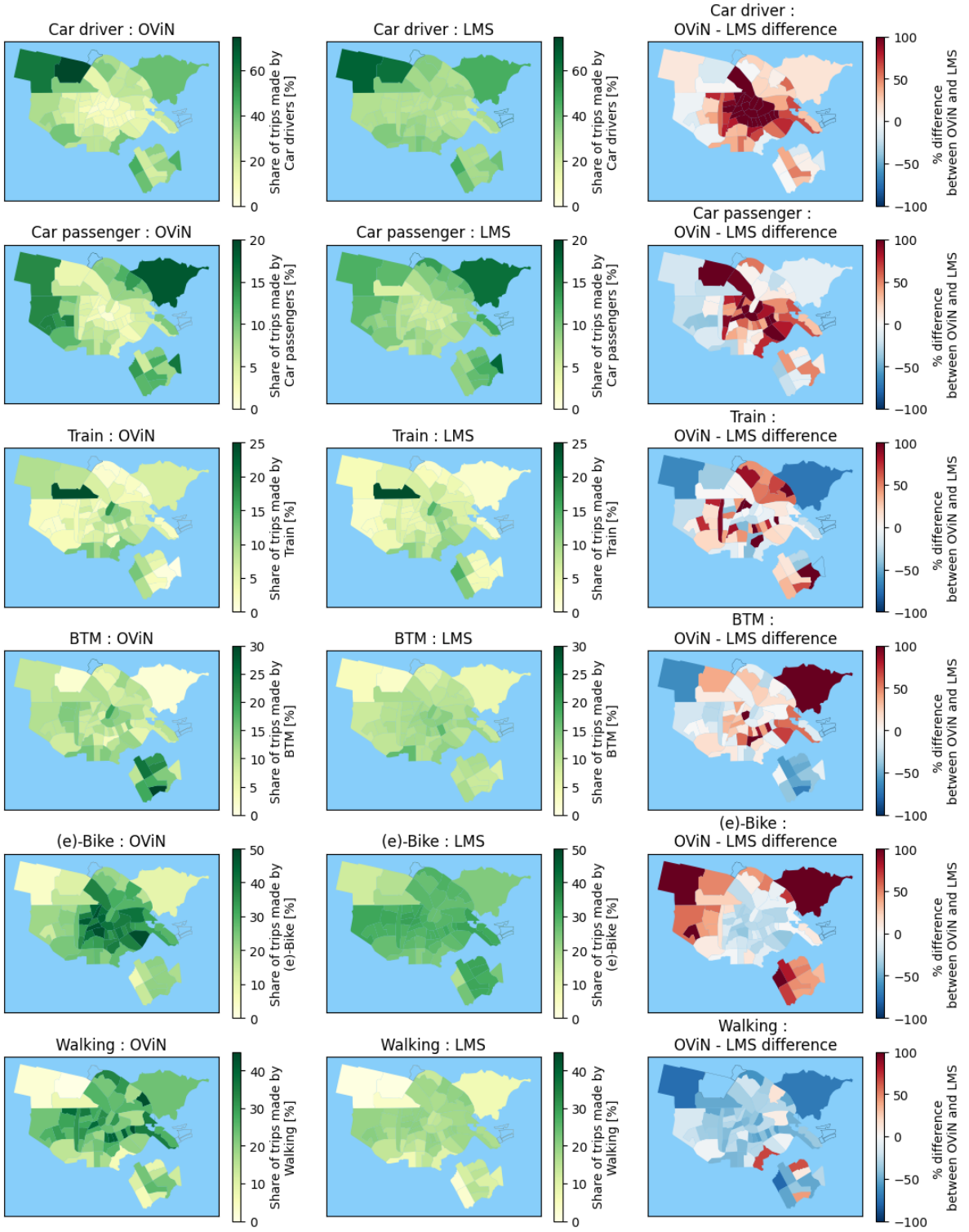

Figure F.1: Left column: Modal split Amsterdam for OViN; Middle column: Modal split Amsterdam for LMS; Right column: Relative difference in modal split for LMS and OViN (difference = (LMS - OViN) / OViN). All modal splits are based on the number of trips departing from a zone. The relative differences are capped at 100 % in this figure. In reality, car driver use in the zones in the city centre are being overestimated by up to 400%. This map is based on the combined OViN dataset for 2013-2017; the LMS OD-matrices (RWS WVL, 2018c) and (RWS WVL, 2020).

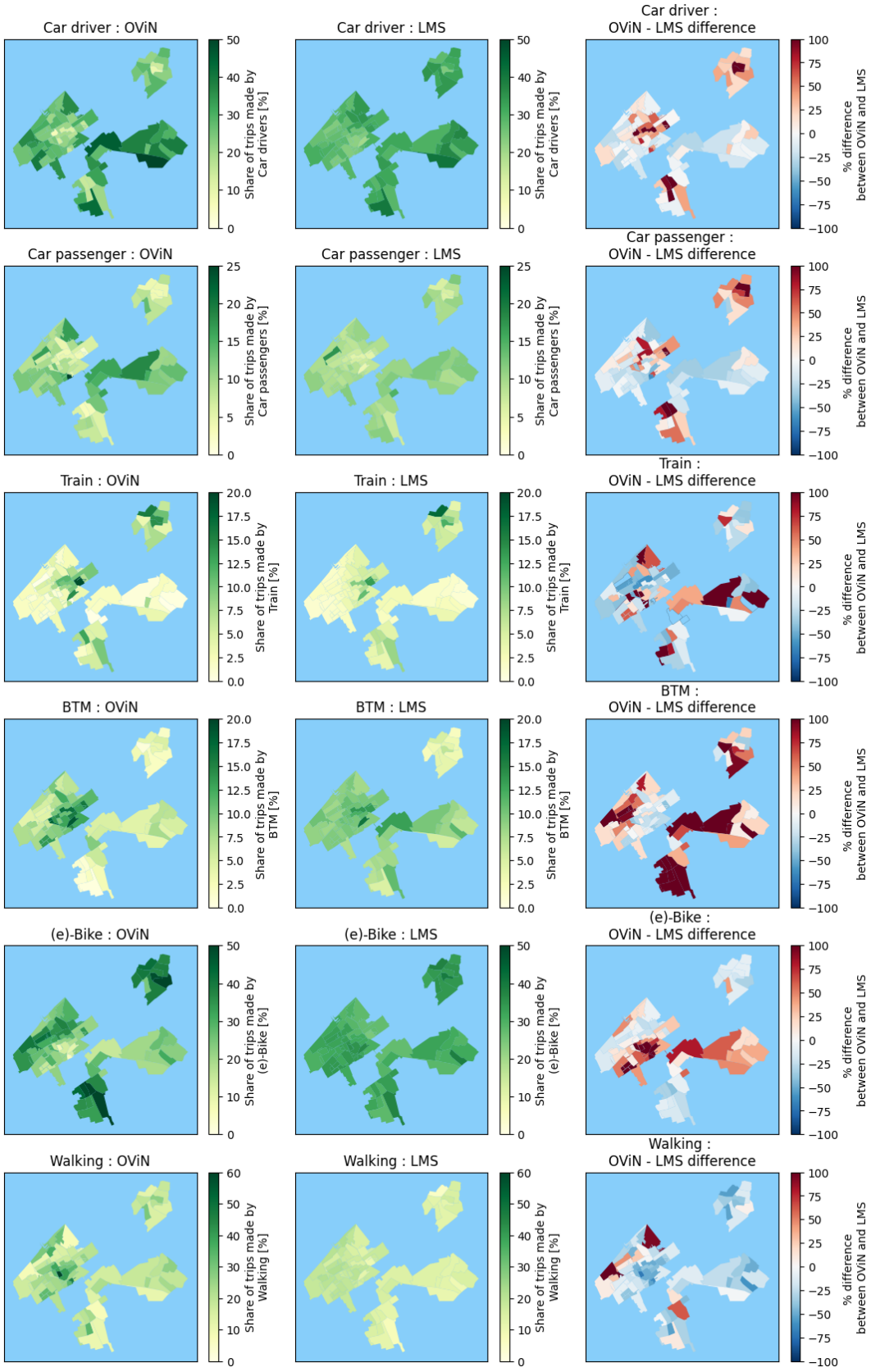

Figure F.2: Left column: Modal split The Hague (top left); Leiden (top right); Zoetermeer (bottom right) and Delft (bottom left) for OViN; Middle column: Modal split The Hague, Leiden, Zoetermeer and Delft for LMS; Right column: Relative difference in modal split for LMS and OViN (difference = (LMS - OViN) / OViN). All modal splits are based on the number of trips departing from a zone. The relative differences are capped at 100 % in this figure. This map is based on the combined OViN dataset for 2013-2017; the LMS OD-matrices (RWS WVL, 2018c) and (RWS WVL, 2020).

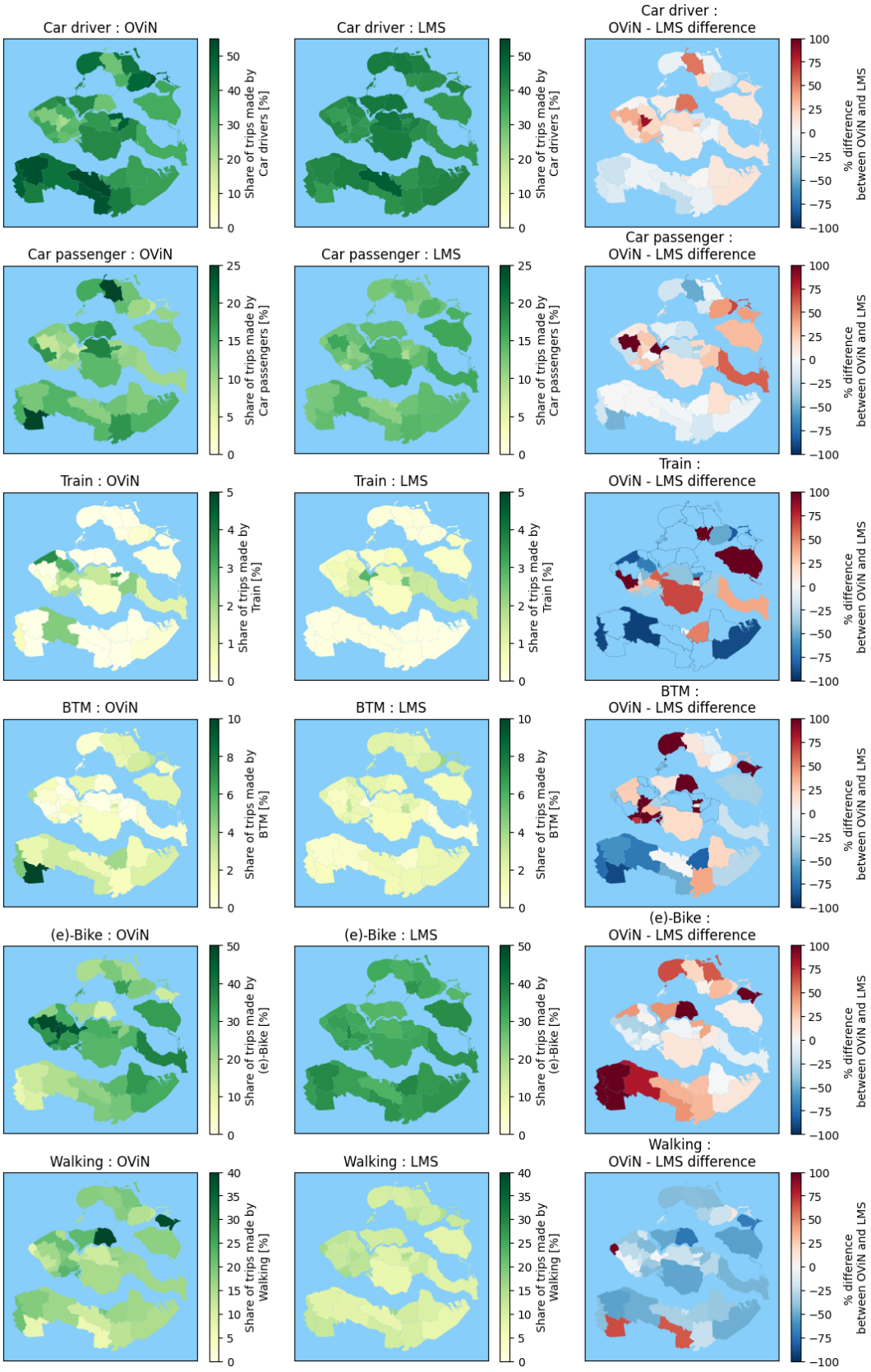

Figure F.3: Left column: Modal split Zeeland for OViN; Middle column: Modal split Zeeland for LMS; Right column: Relative difference in modal split for LMS and OViN (difference = (LMS - OViN) / OViN). All modal splits are based on the number of trips departing from a zone. The relative differences are capped at 100 % in this figure. This map is based on the combined OViN dataset for 2013-2017; the LMS OD-matrices (RWS WVL, 2018c) and (RWS WVL, 2020).

G

Other aspects of travel behaviour

This appendix gives a short overview of the start of an analysis for other aspects of travel behaviour, like travel time, distance and departure time. Due to several reasons, including time and data constraints, this was not finished. See section 5.1 for more details about these limitations in the scope.

G.1. Travel time and distance

An attempt was made to calculate the travel time and distances. For OViN this was very straightforward, because this was already given. For LMS this was given for most modes. For this picture, matrices with average daily travel times and distances were used. However, for example, for train there was no easy way to obtain travel distances. There were separate matrices for station to station and for access and egress, but not for A to B. A lot of time would be needed to get an accurate idea of travel distances. There was a matrix for public transport travel time. It was assumed that this for both train and BTM. The train distances were calculated by fitting a linear regression model for OViN train travel time and distances, and using that to calculate the distance based on the travel time.

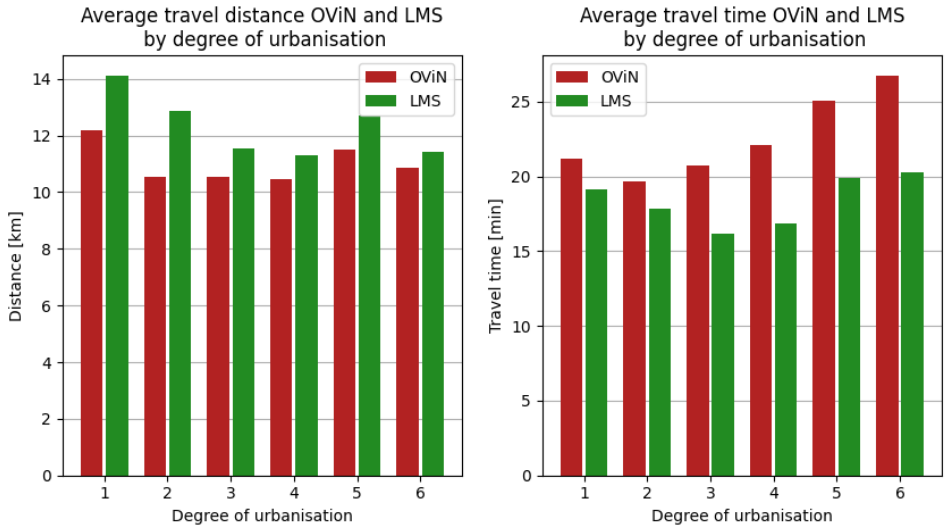

Figure G.1: Differences in average travel time and distance for OViN and LMS

The resulting figure G.1 did not seem logical. All modes are combined here, but the average distance for LMS is longer than for OViN, while the average travel time is less. A possible cause could be because of the overestimation of car travel by LMS and underestimation of walking trips.

For now it is unsure of travel time and distances for OViN and LMS were calculated in the same way, i.e. if these two values are comparable. The results would presumably more logical when separating

travel time and distances per mode or using one way to calculate time and distance for both OViN and LMS. Due to time constraints this was not done.

G.2. Departure time

Figure G.2 shows the share of trips departing at a certain time. It would have been preferable to analyse the modal split for morning peak, for example. This way all people that are departing from home, or arriving to work could be analysed. However, this data was not available for bike and walking, making it not possible to analyse the full modal split. Especially because those modes make up more than half of the trips.

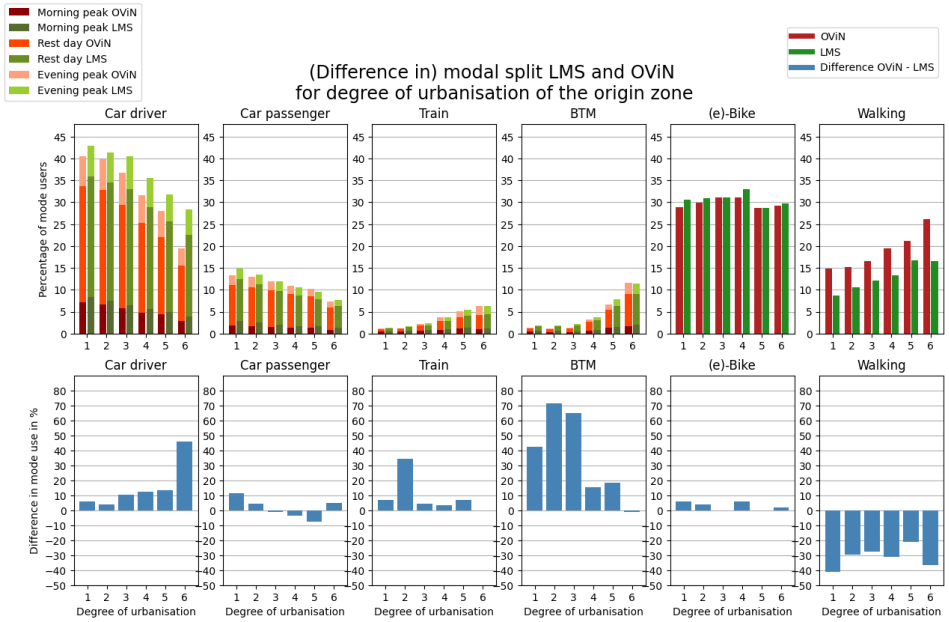

Figure G.2: Share of trips departing during peak hour and rest day.

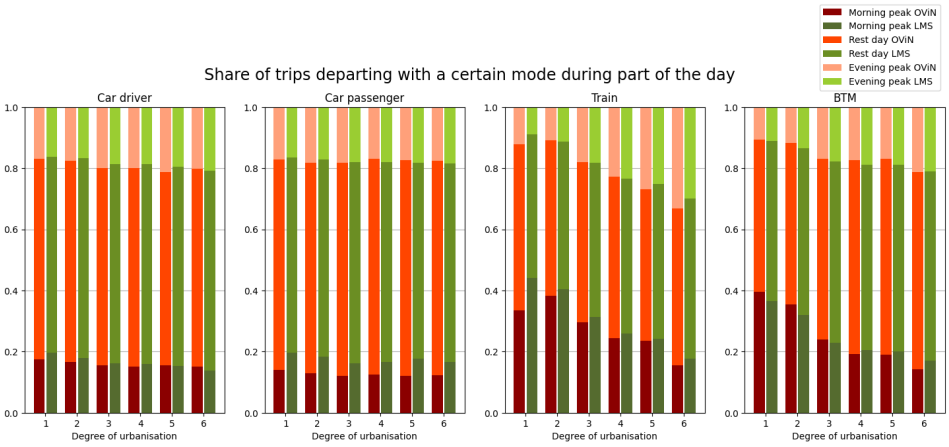

Figure G.3: Relative share of trips departing during peak hour and rest day.

Figure G.3 shows the relative share of trips departing during certain times of day for the car and public transport modes. This figure shows that there are large differences in times of day. For example, in the morning the share of people departing with the train in a region with a low DU is a lot higher than the share of people departing in the evening. This could mean that most people travelling by train to work, work in an area with a high DU. For car travel there seems to be a lot less difference during the day.

Overview of indices used for hierarchical clustering

This appendix gives the different indicators that can be used to help determine the optimal number of clusters. Because of the low number of clusters these scores indicated, they were not used in the final decisions for determining the number of clusters.

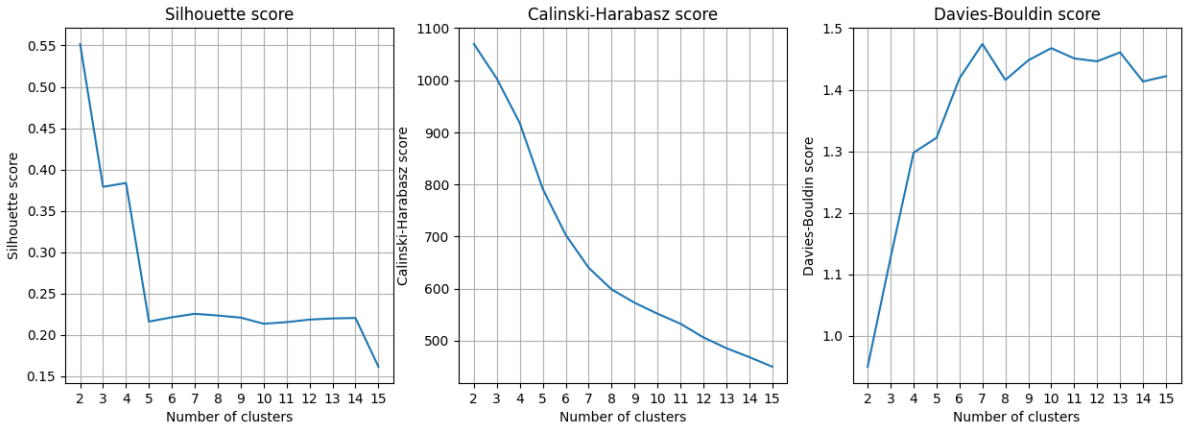

Figure H.1: Indices that help determining the optimal number of clusters for the weighted cluster set. Each index prefers a cluster number of 2. For the variables used to create this cluster set see table and 4.3.

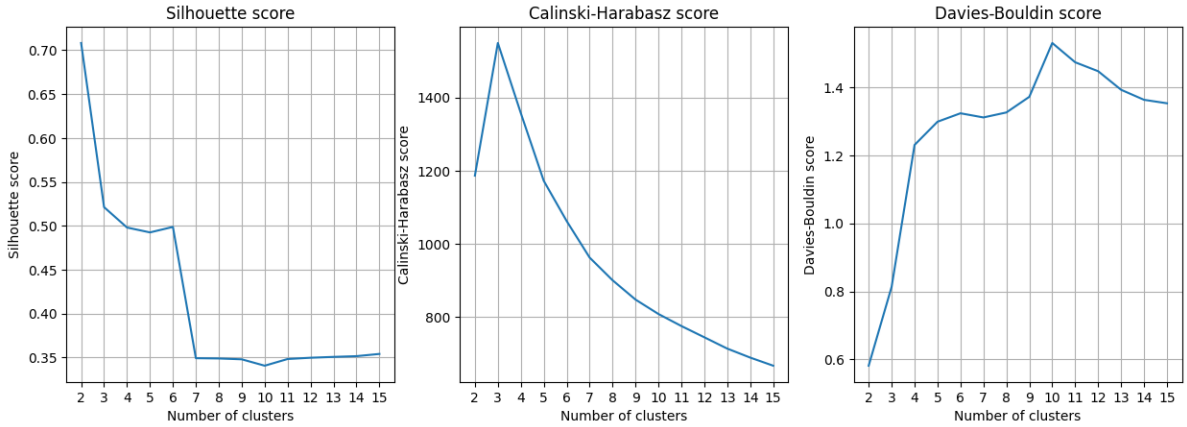

Figure H.2: Indices that help determining the optimal number of clusters for the unweighted cluster set. Each index prefers a cluster number of 2 or 3. For the variables used to create this cluster set see table and 4.5.

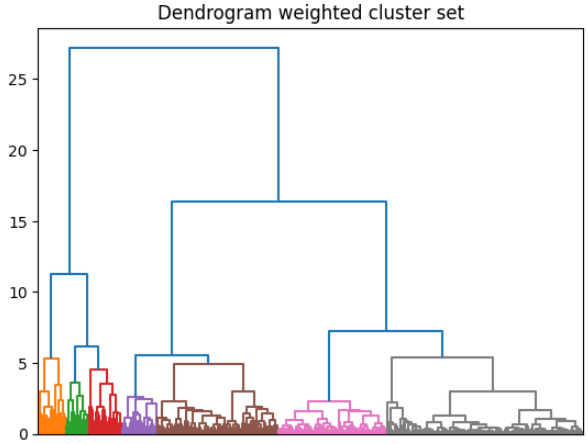

Figure H.3: Dendrogram for the weighted cluster set. Each color shows a different cluster. For the variables used to create this cluster set, see table and 4.3.

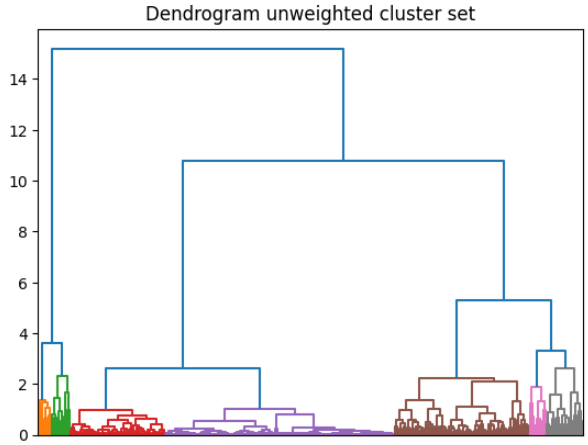

Figure H.4: Dendrogram for the unweighted cluster set. Each color shows a different cluster. For the variables used to create this cluster set, see table and 4.5.

Additional information clustering process

This appendix gives an overview of the process that was done to obtain the final two cluster sets. Table I.2, I.3 and I.4 show alle the different combinations of variables that were tested with some small comments on the quality of the cluster. The number of times a variable appears in the 'Variable' column, is the weight of the variable.

To keep the table compact, all the variables are numbered. They variable is named when they are introduced. Table I.1 gives an overview of all the variables that were used and their numbers.

Table I.1: Overview of all the variables used to create the clusters.

Number	Variable name	Description	D-variable
0	Pop_dens	Population density [people/ha]	Density
1	Surrounding_pop_dens	Population density of all zones in a radius of 3 km [people/ha]	Density
2	DU	Degree of urbanisation	Density
3	Job_dens	Job density [jobs/ha]	Density
4	Surrounding_job_dens	Job density of all zones in a radius of 3 km [jobs/ha]	Density
5	Residential	Ratio of landuse used for residential	Diversity
6	Services	Ratio of landuse used for services	Diversity
7	Industrial	Ratio of landuse used for industrial	Diversity
8	Nature	Ratio of landuse used for nature	Diversity
9	Agricultural	Ratio of landuse used for agriculture	Diversity
10	Infra	Ratio of landuse used for infrastructure	Diversity
11	Entropy	Entropy measure	Diversity
12	Special	National specialisation index	Diversity
13	House_45_less	Ratio of houses built before 1945	Diversity
14	House_45_75	Ratio of houses built between 1945 and 1975	Diversity
15	House_75_05	Ratio of houses built between 1975 and 2005	Diversity
16	House_05_more	Ratio of houses built after 2005	Diversity
17	Job-workers ratio	Working population / number of jobs	Diversity
18	Road_density	Length of road per area [km/ km2]	Design
19	Road_width	Average road width [m]	Design
20	Road_parking	Area used for parking next to roads / total population [m2/person]	Demand management
21	Bike_walk_percentage	Share of road meant for walking or cycling	Design
22	Dist_to_center	Average distance to city centre [km]	Destination accessibility
23	Dist_food	Average minimum distance to several food related locations [km]	Destination accessibility
24	Dist_commercial	Average minimumdistance to several commercial related locations [km]	Destination accessibility
25	Dist_health	Average minimum distance to several health related locations [km]	Destination accessibility
26	Dist_recreation	Average minimum distance to several recreation locations [km]	Destination accessibility
27	Dist_education	Average minimum distance to several education related locations [km]	Destination accessibility
28	Dist_point_of_interest	Average minimum distance to several points of interest [km]	Destination accessibility
29	Parking_fare	Average parking fee [euros]	Demand management
30	Distance_station	Distance to closest train station [km]	Distance to transit
31	Distance_ic_station	Distance to closest intercity train station [km]	Distance to transit
32	Freq_station	Train frequency of closest train station [trains/hour]	Distance to transit
33	Freq_ic_station	Train frequency of closest intercity train station [trains/hour]	Distance to transit
34	Distance_btm	Average distance to a btm stop [km]	Distance to transit
35	Distance_bus	Average distance to a bus stop [km]	Distance to transit
36	Distance_metro	Average distance to a metro stop [km]	Distance to transit
37	Distance_tram	Average distance to a tram stop [km]	Distance to transit
38	Btm_lines	Number of different btm lines of closest stops	Distance to transit
39	Bus_lines	Number of different bu lines of closest stops	Distance to transit
40	Metro_lines	Number of different metro lines of closest stops	Distance to transit
41	Tram_lines	Number of different tram lines of closest stops	Distance to transit
42 (new variable)	btm_stops	Number of btm stops within a certain radius	Distance to transit
43 (new variable)	Bus_stops	Number of bus stops within a certain radius	Distance to transit
44 (new variable)	Metro_stops	Number of metro stops within a certain radius	Distance to transit
45 (new variable)	Tram_stops	Number of tram stops within a certain radius	Distance to transit
46 (previously 42)	Distance_TM	Average distance to a tram or metro stop [km]	Distance to transit
47 (previously 43)	Nature_Agri	Ratio of landuse used for agriculture or nature	Diversity
48	TM_stops	Number of tram and metro stops within a certain radius	Distance to transit

Table I.2: Iterations that were done to obtain the final cluster sets, part 1. The 'Cluster set' column indicates to which cluster set this iteration belongs. U means unweighted cluster set and W means weighted cluster set.

Iteration	Cluster set	Variables	New variable	Keep?	Comments
1	both	4	4: Surrounding_job_dens	Yes	
2	both	4, 0	0: Pop_dens	Yes	
3	both	4, 0, 1	1: surrounding_pop_dens	Yes	Adding this variable makes the clusters a lot more interesting. Especially with 7 clusters.
4	U	4, 0, 1, 29	29: Parking_fare	Yes	It is interesting, but makes a lot of small clusters. This means that there is a lot of overfitting.
5	W	4, 0, 1, 29, 29, 29		Yes	The overfitting becomes less when introducing the weights. In this case Demand management and Density both have an equal weight.
6	U	4, 0, 1, 29, 18	18: Road_density	Yes	The clusters are fairly good, even with only 5 clusters.
7	W	4, 0, 1, 29, 29, 29, 18, 18, 18		Yes	Road density also looks good when adding weights. However, there is more overfitting
8	U	4, 0, 1, 29, 18, 2	2: DU	No	Clusters are sorted primarily based on the DU. Not interesting for this thesis
9	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 2		No	This looks better, because the DU becomes less important. However, it has been chosen to not use the DU while making the clusters.
10	U	4, 0, 1, 29, 18, 3	3: Job_density	No	There is a lot of overfitting. The less urban areas are all places in the same cluster.
11	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3		Yes	This looks better. There is still overfitting, but more interesting variation. It has been decided to include this variable.
12	U	4, 0, 1, 29, 18, 42	42: Distance_TM	Yes	A lot of variation in clusters, especially in the Randstad, which has a lot of TM. The rest of NL is underfitted. Looks better when using 7 or 8 clusters.
13	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42,42, 42		Yes	It looks fairly good. There is not a lot of variation in rural areas, 2 suburb clusters. When using 7 clusters, there is an improvement in rural areas.
14	U	4, 0, 1, 29, 18, 42, 37	37: Distance_tram	No	
15	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 37, 37, 36, 36	36: Distance_metro	No	Replace 42 for 36 and 37. This might be interesting for a smaller model (NRM?), but the clusters are too much focused on different PT networks.
16	U	4, 0, 1, 29, 18, 42, 5	5: Residential	Yes	With this variable, smaller cities pop out more. 6 clusters gives better results than 7.
17	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 42, 42, 5, 5, 5, 5		Maybe	The clusters do not seem to become significantly better. However, 5 might still be an interesting variable when more diversity variables are used.
18	U	4, 0, 1, 29, 18, 42, 5, 12	12: Special	Maybe	No clear improvements. But might be interesting to add again later.
19	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 42, 42, 5, 5, 12, 12		Maybe	The medium-sized cities are kind of forgotten when using this variable.
20	U	4, 0, 1, 29, 18, 42, 5, 6	6: Services	Yes	Clusters look a bit better. Though could be removed later, perhaps.
21	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 42, 42, 5, 5, 6, 6		Yes	Seems to be able to identify more cities and villages.
22	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 42, 42, 5, 6, 12, 12		Yes	More different cities are identified, with clear medium-sized cities with high bike use. The centra of the large cities are more fragmented. When using 7 clusters, rural areas show more distinctions.
23	U	4, 0, 1, 29, 18, 42, 5, 6, 9	9: Agricultural	Maybe	This variable gives 2 clear clusters in rural areas. It might overfit a bit
24	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 42, 42, 5, 6, 9, 12		Maybe	Better. More variation in rural areas, with less overfitting. However, the Veluwe and the Dunes pop out. Not sure if that is desirable.
25	U	4, 0, 1, 29, 18, 42, 5, 6, 43	43: Nature_Agri	No	Adding 43 as an attempt to solve the "Veluwe problem". This gives a lot of fragmentation, so it is not a succes.
26	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 42, 42, 5, 6, 43, 12		Maybe	Seems okay with 7 clusters.
27	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 42, 42, 5, 6, 43, 9		Maybe	Looks better. More variation in rural areas. However, still a strong bias for the Veluwe.
28	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 42, 42, 5, 6, 8, 9	8: Nature	No	Too much focus on rural areas. Medium-sized cities disappear
29	W			Maybe	This is a bit better.
30	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 42, 42, 12, 6, 43, 43		Yes	Using 12 instead of 5 gives slightly better results. Use this combination for now.
31	U	4, 0, 1, 29, 18, 42, 12, 6, 43		Yes	Looks good with 7 clusters.
32	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 23	23: Dist_food	No	Adding this variable places too much focus on the Randstad.
33	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28	28: Dist_point_of_interest	Yes	Shows clear medium-sized cities with 7 clusters. However, there is a lot of fragmentation in the Randstad.
34	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 42, 42, 12, 6, 43, 43, 28, 28, 28, 28		Yes	Looks good with 7 clusters. However, the cluster size of the large city centres is very small.
35	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 27	27: Dist_education	No	Interesting, but medium-sized cities disappear. Rural areas look better. See if the latter effect can be maintained.
36	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 42, 42, 12, 6, 43, 43, 28, 28, 27, 27		Maybe	Increases quality clusters in rural areas. However, medium-sized cities disappear.
37	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 32	32: Freq_station	No	Gives more differentiation in Randstad. It does not seem like an improvement.
38	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 32, 32, 12, 6, 43, 43, 28, 28, 28, 28		Yes	Seems to be the best so far. There is a good balance between meidum-sized cities and rural. Medium-sized city cluster is a bit small. However, Delft is in the same cluster as Zoetermeer.
39	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 32, 32, 12, 6, 43, 43, 28, 28, 28, 27		No	Interesting, but iteration 38 seemed better. There is more overfitting in the Randstad. It does have potential.
40	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 26	26: Dist_recreation	No	Medium-sized cities disappear, randstad is overfitted.
41	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 32, 32, 12, 6, 43, 43, 28, 28, 28, 26		No	No improvements observed.
42	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 13	13: House_45_less	No	Medium-sized cities disappear, randstad is overfitted.
43	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 32, 32, 12, 6, 13, 43, 28, 28, 28, 28		Maybe	This combination has potential. However, variation in rural areas disappears.
44	W	4, 0, 1, 1, 29, 29, 29, 29, 29, 18, 18, 18, 18, 18, 3, 42, 42, 42, 32, 32, 12, 6, 13, 43, 43, 28, 28, 28, 28, 28		No	Only Delft seems to become noticable better based in this clusterset. It seems like needlessly complicating clusters (different weights D-variables).
45	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 25	25: Dist_health	No	No improvements observed.
46	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 32, 32, 12, 6, 13, 43, 28, 28, 28, 25		No	No improvements observed.
47	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 24	24: Dist_commercial	No	No improvements observed.
48	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 32, 32, 12, 6, 13, 43, 28, 28, 28, 24		No	No improvements observed.
49	W	4, 0, 1, 1, 29, 29, 29, 29, 29, 18, 18, 18, 18, 18, 3, 42, 42, 42, 32, 32, 12, 6, 43, 43, 43, 27, 26, 25, 24, 23		No	Tried using different distance variables instead of 28. However, medium-sized cities disappear and Randstad is overfitted. It appears that using one measure for Destination accessibility gives better results.
50	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 31	31: Distance_ic_station	No	New rural clusters that appear show barely any differences in travel behaviour. There is a lot of overfitting with 8 clusters.

Table I.3: Iterations that were done to obtain the final cluster sets, part 2. The 'Cluster set' column indicates to which cluster set this iteration belongs. U means unweighted cluster set and W means weighted cluster set.

Iteration	Cluster set	Variables	New variable	Keep?	Comments
51	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 32, 31, 12, 6, 43, 43, 28, 28, 28, 28		No	To little differentiation of rural areas and too much focus on the Randstad.
52	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 31, 32, 31, 12, 6, 43, 43, 28, 28, 28, 28		No	See comment previous iteration.
53	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 10	10: Infra	No	Overfitting in the Randstad.
54	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 32, 32, 12, 6, 10, 43, 28, 28, 28, 28		No	No improvements observed.
55	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 34	34: Distance_btm	No	Overfitting in the Randstad.
56	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 34, 32, 32, 12, 6, 43, 43, 28, 28, 28, 28		No	It does have some potential with interesting clusters. So perhaps this variable can be included in a different way.
57	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 34, 32, 12, 6, 43, 43, 28, 28, 28, 28		No	This clusterset performs worse than the previous iteration.
58	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 33	33: Freq_ic_station	No	The clusters in urban areas seem strange. Clusters in rural areas are more interesting.
59	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 32, 33, 12, 6, 43, 43, 28, 28, 28, 28		No	No improvements observed.
60	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 21	21: Bike_walk_percentage	No	No improvements observed.
61	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 21, 21, 3, 42, 42, 32, 32, 12, 6, 43, 43, 28, 28, 28, 28		No	No improvements observed.
62	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 35	35: Distance_bus	No	It could possibly work with 5 clusters, for a very simple cluster set.
63	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 35, 42, 32, 32, 12, 6, 43, 43, 28, 28, 28, 28		No	Interesting with 7 clusters, but no clear improvements are observed.
64	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 35, 42, 42, 32, 12, 6, 43, 43, 28, 28, 28, 28		No	Differentiation in rural areas disappears.
65	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 35, 35, 42, 32, 12, 6, 43, 43, 28, 28, 28, 28		No	Differentiation in rural areas disappears.
66	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 22	22: Dist_to_center	No	No improvements observed.
67	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 22		Maybe	This clusterset seems okay, but no clear improvements.
68	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 32, 32, 12, 6, 43, 43, 22, 22, 22, 22		No	Differentiation in rural areas disappears.
69	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 32, 32, 12, 6, 43, 43, 28, 28, 22, 22		No	Similar to iteration 38, so adding an additional Destination accessibility variable only makes the clusters more complicated.
70	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 30	30: Distance_station	Maybe	There is a decrease in the quality of the medium-sized city cluster. However, interesting rural patterns appear.
71	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 32, 30, 12, 6, 43, 43, 28, 28, 28, 28		No	Overfitting in the Randstad.
72	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 32, 32, 30, 12, 6, 43, 43, 28, 28, 28, 28		Maybe	When using 7 clusters, an additional rural cluster appears with very low train use. With 6 clusters also okay, but no improvemtn over iteration 38.
73	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 14	14: House_45_75	No	No improvements observed.
74	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 20	20: Road_parking	No	No improvements observed.
75	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 11	11: Entropy	No	Medium-sized cities disappear.
76	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 32, 32, 11, 6, 12, 43, 28, 28, 28, 28		Maybe	This clusterset looks very interesting with 7 or 8 clusters.
77	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 17	17: Job-workers ratio	No	No improvements observed.
78	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 7	7: Industrial	No	Overfitting in the Randstad.
79	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 19	19: Road_width	No	Overfitting in the Randstad.
80	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 15	15: House_75_05	No	No improvements observed.
81	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 28, 16	16: House_05_more	No	Okay with 7 clusters, bu no clear improvements observed. Overfitting Randstad.
82	U	4, 0, 1, 29, 18, 42, 12, 6, 43, 13		No	Looks interesting, but little differentiation rural zones.
General problem: all zones that have a tram or metro stop within a certain radius are automatically put into a separate cluster. This gives that Randstad clusters are automatically put in different clusters than the rest of the country.					
83	U	4, 0, 1, 29, 18, 12, 6, 43		No	Without a variable for tram/metro, the density variables become too important and other interesting trends disappear.
84	U	4, 0, 1, 29, 18, 42, 6, 43		No	No improvements observed.
85	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 42, 32, 32, 12, 6, 43, 43, 28, 28, 28, 28		Maybe	This clusterset seems to be a small improvement. 2 clear more rural clusters.
86	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 36, 37, 32, 32, 13, 6, 43, 43, 28, 28, 28, 28		No	No improvements observed.
87	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 36, 37, 32, 32, 13, 6, 13, 43, 28, 28, 28, 28		Maybe	Looks interesting. However, a way must be found to better distinguish between cities.
88	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 36, 30, 32, 32, 13, 6, 13, 43, 28, 28, 28, 28		Yes	Looks interesting. The distinction between Randstad and the rest of NL becomes less obvious (which is positive).
89	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 30, 32, 32, 13, 6, 13, 43, 28, 28, 28, 28		Maybe	Small improvement in largest cities
90	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 42, 31, 32, 32, 13, 6, 13, 43, 28, 28, 28, 28		Maybe	No improvements observed.
91	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 36, 31, 32, 32, 13, 6, 13, 43, 28, 28, 28, 28		No	The clustering for The Hague and Utrect clearly becomes worse.
Note: new improved variables are added. Variable 42 is now 46 and 43 is 47					
92	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 32, 32, 13, 6, 13, 47, 28, 28, 28, 28	44: Metro_stops 45: Tram_stops	Yes	New variables seem to improve the clusters in the Randstad.
93	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 32, 32, 12, 6, 12, 47, 28, 28, 28, 28		No	No improvements observed.
94	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 32, 32, 12, 6, 13, 47, 28, 28, 28, 28		Maybe	Slightly better than the previous iteration. However, the medium-sized cities could be highlighted better.
95	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 12, 6, 13, 47, 28, 28, 28, 28	42: Bus_stops	Maybe	It looks fairly okay, but medium-sized cities can still be improved.

Table I.4: Iterations that were done to obtain the final cluster sets, part 3. The 'Cluster set' column indicates to which cluster set this iteration belongs. U means unweighted cluster set and W means weighted cluster set.

Iteration	Cluster set	Variables	New variable	Keep?	Comments
96	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 13, 6, 13, 47, 28, 28, 28, 28		Yes	This seems like the best iteration so far. Medium-sized cities could be improved.
97	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 43, 13, 6, 13, 47, 28, 28, 28, 28		No	No improvements observed.
98	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 13, 6, 13, 47, 28, 28, 22, 22		No	Medium-sized cities disappear.
99	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 13, 13, 13, 47, 28, 28, 28, 28		No	No improvements observed.
100	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 6, 6, 13, 47, 28, 28, 28, 28		No	Weird distinction between different cities.
101	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 31, 13, 6, 13, 47, 28, 28, 28, 28		No	Medium-sized cities disappear.
102	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 32, 13, 6, 13, 47, 28, 28, 28, 28		Maybe	A slight improvement can be observed compared to previous iterations.
103	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 13, 12, 13, 47, 28, 28, 28, 28		No	No improvements observed.
104	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 13, 6, 13, 47, 28, 28, 28, 28, 28, 28		No	No improvements observed.
105	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 13, 6, 13, 47, 28, 28, 28, 28,13		No	The borders of NL are highlighted too much in a separate cluster.
106	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 13, 5, 13, 47, 28, 28, 28, 28		Maybe	A slight improvement can be observed compared to previous iterations.
107	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 6, 5, 13, 47, 28, 28, 28, 28		No	Medium-sized cities disappear.
108	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 6, 5, 13, 13, 28, 28, 28, 28		Maybe	This iteration looks okay. Especially with 7 clusters.
109	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 4, 45, 44, 43, 32, 6, 5, 13, 13, 28, 28, 28, 28		No	No improvements observed.
110	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 43, 6, 5, 13, 13, 28, 28, 28, 28		Maybe	This clusterset looks good with 7 clusters.
111	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 42, 6, 5, 13, 13, 28, 28, 28, 28	42: btm_stops	Maybe	Even more improvements at 7 clusters.
112	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 42, 6, 5, 13, 13, 28, 28, 28, 22		No	No improvements observed.
A mistake was discovered in the variables 42-45. These are now updated.					
113	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 13, 5, 13, 47, 28, 28, 28, 28		No	Decrease in quality after updating variables
114	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 13, 5, 13, 47, 28, 28, 28, 28		No	Decrease in quality after updating variables
115	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 13, 47, 13,6, 28, 28, 28, 28		No	Decrease in quality after updating variables
116	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 13, 5, 13,6, 28, 28, 28, 28		Yes	This clusterset looks good at 7 clusters.
117	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 48, 13, 5, 13,6, 28, 28, 28, 28	48: TM_stops	No	No improvements observed.
118	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 42, 13, 5, 13,6, 28, 28, 28, 28		No	The clusters seem unlogical.
119	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 43, 13, 5, 13,6, 28, 28, 28, 28		No	Medium-sized cities disappear and the Randstad is overfitted.
120	W	4, 0, 0, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 13, 5, 13,6, 28, 28, 28, 28		No	Medium-sized cities disappear.
121	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 32, 13, 5, 13,5, 28, 28, 28, 28		No	Large urban city centres are overfitted.
122	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 45, 13, 5, 13,6, 28, 28, 28, 28		No	No improvements observed.
123	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 45, 44, 43, 44, 13, 5, 13,6, 28, 28, 28, 28		Maybe	Small improvements can be observed compared to previous iterations.
124	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 48, 48, 43, 32, 13, 5, 13,6, 28, 28, 28, 28		No	No improvements observed.
125	W	4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 48, 48, 43, 32, 13, 6, 13,6, 28, 28, 28, 28		Yes	The cluster for the large urban areas is very small. Medium-size city cluster looks good. Best with 7 clusters.
126	W	**4, 0, 1, 29, 29, 29, 29, 18, 18, 18, 18, 3, 48, 48, 43, 43, 13, 6, 13,6, 28, 28, 28, 28**		Yes	Final version! Clear cluster for medium-sized cities. 6 clusters are good, but 7 clusters shows even more interesting patterns. (This will be further elaborated in the thesis.)
127	U	4, 0, 1, 29, 18, 48, 6, 47		No	With improved variable the cluster does not look good.
128	U	**4, 0, 1, 29, 18, 48, 6**		Yes	Final version! Nice clusters. With 7 clusters smaller cities are highlighted better.

J

Additional information PSM

This appendix gives additional information about the PSM. First the tables J.1, J.2 and J.3 show the average demographic characteristics of each cluster before matching. The tables confirm that the differences in demographics can be large between clusters. For example, the average age in the rural cluster (cluster 0) of the weighted cluster set is almost 3 years higher than the cluster of the large urban centres (cluster 1).

Table J.1: An overview of the average demographics for each cluster from the weighted cluster set before matching. The reference variable for each category is written in *italic* (e.g. 2 parent household in the household type category). The demographic characteristics are based on the combined OViN dataset from 2013-2017.

Demographic characteristic	**Cluster**						
	0	**1**	**2**	**3**	**4**	**5**	**6**
Age	39.24	36.42	37.15	38.40	39.11	38.21	39.53
Gender	0.54	0.52	0.53	0.53	0.53	0.52	0.53
Income footnote	5.58	5.13	5.39	5.44	5.71	5.26	5.57
Household size	3.19	2.68	2.67	3.03	3.10	2.89	2.91
1 person household	0.11	0.26	0.26	0.14	0.12	0.19	0.17
2+ person household	0.27	0.26	0.28	0.25	0.26	0.24	0.27
1 parent household	0.06	0.11	0.07	0.09	0.07	0.12	0.08
2 parent household	*0.56*	*0.38*	*0.40*	*0.52*	*0.55*	*0.45*	*0.58*
Part time workers	0.16	0.11	0.15	0.15	0.15	0.11	0.14
Full time workers	0.30	0.42	0.37	0.31	0.31	0.37	0.33
Students	0.20	0.20	0.24	0.21	0.20	0.21	0.21
Other participation	*0.34*	*0.26*	*0.24*	*0.33*	*0.33*	*0.31*	*0.32*
Primary education or less	0.05	0.06	0.05	0.06	0.06	0.06	0.06
lbo or vmbo education	0.18	0.09	0.10	0.15	0.16	0.13	0.14
mbo, havo or vwo education	0.33	0.26	0.31	0.32	0.32	0.28	0.31
hbo or wo	*0.22*	*0.45*	*0.42*	*0.28*	*0.27*	*0.35*	*0.33*
Other education	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Younger than 15 years	0.20	0.13	0.11	0.18	0.19	0.16	0.16
Number of cars per household	1.56	0.81	1.05	1.28	1.43	1.02	1.26
Drivers licence	0.71	0.66	0.73	0.68	0.71	0.65	0.71
Student OV	0.03	0.08	0.12	0.04	0.04	0.06	0.05

The figures J.1, J.2 and J.3 show the PS for each cluster pair for both cluster sets and the DU. These graphs shows that there is sufficient overlap between the demographics of each clusters to make matches. These graphs also show how some clusters have many more data points than others. The tables J.4, J.5 and J.6 show all ATE, OBE and ATE-OBE ratio values for each cluster pair and each DU pair. All values where $p > 0.05$ after the t-test are removed. The figures J.4 and J.5 show the standard mean deviation values for each cluster pair and each DU pair before and after matching.

Table J.2: An overview of the average demographics for each cluster from the unweighted cluster set before matching. The reference variable for each category is written in *italic* (e.g. 2 parent household in the household type category). The demographic characteristics are based on the combined OViN dataset from 2013-2017.

Demographic characteristic	Cluster						
	0	1	2	3	4	5	6
Age	36.65	36.81	38.34	38.59	35.81	39.38	38.94
Gender	0.53	0.52	0.52	0.53	0.53	0.54	0.53
Income footnote	5.27	5.27	5.45	5.46	4.67	5.62	5.68
Household size	2.84	2.61	2.61	3.00	2.92	3.16	3.08
1 person household	0.21	0.27	0.26	0.15	0.20	0.11	0.13
2+ person household	0.24	0.27	0.29	0.25	0.21	0.27	0.26
1 parent household	0.11	0.10	0.07	0.09	0.14	0.06	0.07
2 parent household	*0.44*	*0.36*	*0.38*	*0.51*	*0.44*	*0.56*	*0.54*
Part time workers	0.13	0.12	0.14	0.15	0.10	0.16	0.15
Full time workers	0.36	0.44	0.38	0.32	0.34	0.30	0.32
Students	0.24	0.20	0.23	0.21	0.23	0.20	0.21
Other participation	*0.28*	*0.24*	*0.24*	*0.33*	*0.34*	*0.34*	*0.33*
Primary education or less	0.06	0.05	0.05	0.06	0.08	0.05	0.06
lbo or vmbo education	0.11	0.08	0.10	0.15	0.14	0.18	0.15
mbo, havo or vwo education	0.28	0.25	0.33	0.32	0.28	0.33	0.31
hbo or wo	*0.38*	*0.49*	*0.43*	*0.29*	*0.31*	*0.23*	*0.28*
Other education	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Younger than 15 years	0.16	0.12	0.09	0.18	0.18	0.20	0.19
Number of cars per household	1.00	0.78	1.08	1.27	0.93	1.55	1.39
Drivers licence	0.66	0.68	0.76	0.68	0.61	0.71	0.70
Student OV	0.09	0.08	0.13	0.05	0.07	0.03	0.04

Table J.3: An overview of the average demographics for each DU before matching. The reference variable for each category is written in *italic* (e.g. 2 parent household in the household type category). The demographic characteristics are based on the combined OViN dataset from 2013-2017.

Demographic characteristic	Degree of urbanisation					
	1	2	3	4	5	6
Age	39.42	39.75	39.07	38.02	37.04	37.03
Gender	0.53	0.54	0.53	0.53	0.54	0.51
Income footnote	5.52	5.65	5.59	5.46	5.50	5.20
Household size	3.19	3.12	3.06	2.95	2.92	2.73
1 person household	0.11	0.11	0.13	0.17	0.19	0.24
2+ person household	0.28	0.28	0.26	0.25	0.24	0.25
1 parent household	0.05	0.06	0.07	0.09	0.1	0.11
2 parent household	*0.56*	*0.55*	*0.53*	*0.49*	*0.47*	*0.40*
Part time workers	0.16	0.16	0.15	0.14	0.13	0.11
Full time workers	0.30	0.29	0.31	0.33	0.36	0.41
Students	0.20	0.20	0.21	0.22	0.23	0.20
Other participation	*0.34*	*0.35*	*0.33*	*0.31*	*0.28*	*0.27*
Primary education or less	0.06	0.06	0.06	0.06	0.06	0.06
lbo or vmbo education	0.18	0.18	0.16	0.14	0.12	0.10
mbo, havo or vwo education	0.34	0.33	0.32	0.32	0.29	0.26
hbo or wo	*0.22*	*0.24*	*0.27*	*0.31*	*0.36*	*0.43*
Other education	0.01	0.01	0.01	0.01	0.01	0.01
Younger than 15 years	0.2	0.19	0.18	0.17	0.16	0.14
Number of cars per household	1.58	1.52	1.39	1.22	1.09	0.87
Drivers licence	0.72	0.72	0.70	0.69	0.66	0.66
Student OV	0.03	0.03	0.04	0.07	0.08	0.07

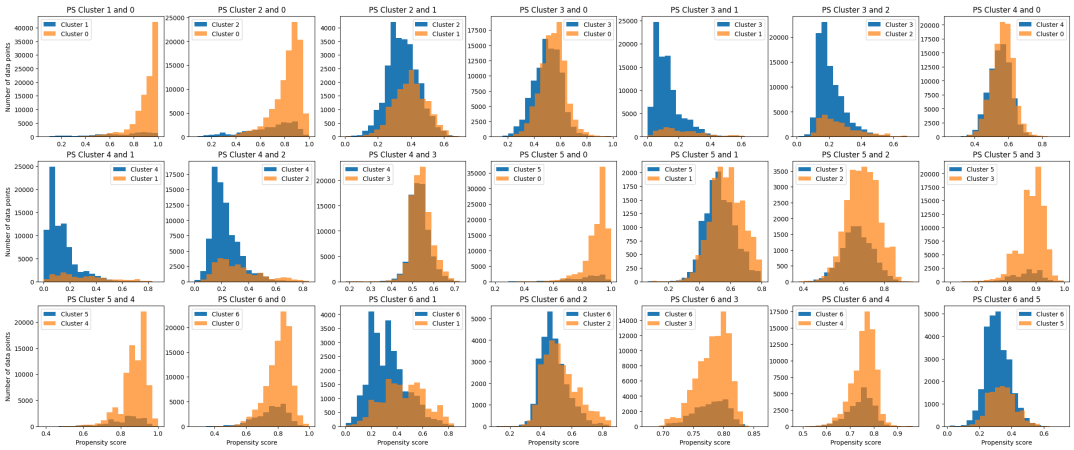

Figure J.1: Propensity scores for the weighted cluster set for all cluster pairs. There is sufficient overlap to perform PSM. The demographic characteristics used to calculate the propensity scores are based on the combined OViN dataset from 2013-2017.

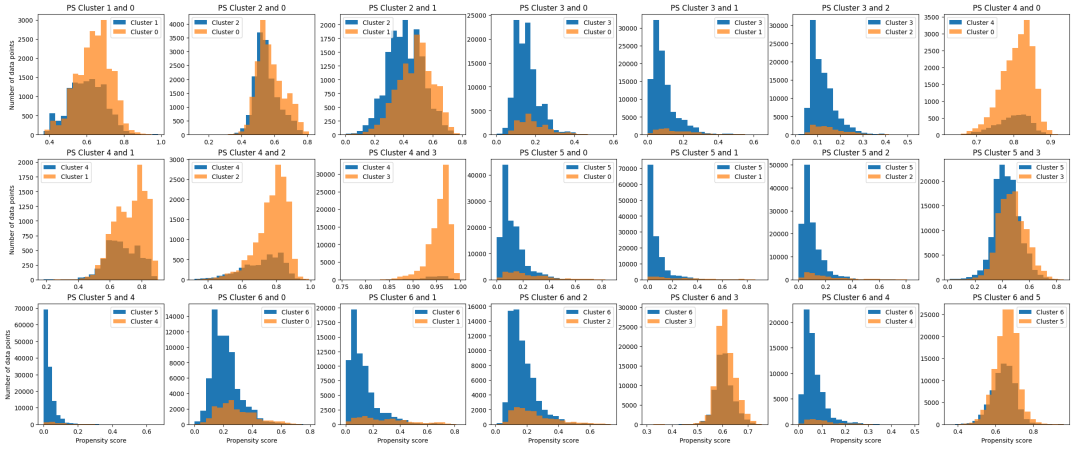

Figure J.2: Propensity scores for the unweighted cluster set for all cluster pairs. There is sufficient overlap to perform PSM. The demographic characteristics used to calculate the propensity scores are based on the combined OViN dataset from 2013-2017.

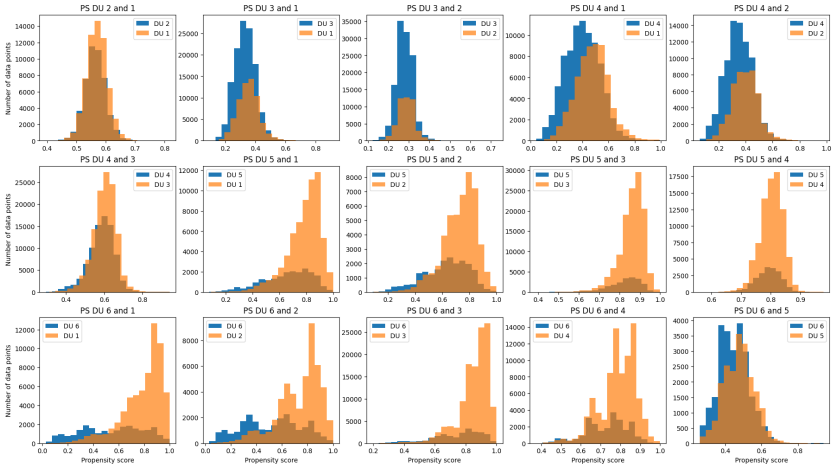

Figure J.3: Propensity scores for each degree of urbanisation pair. There is sufficient overlap to perform PSM. The demographic characteristics used to calculate the propensity scores are based on the combined OViN dataset from 2013-2017.

Table J.4: ATE, OBE and the ATE OBE ratio for the weighted cluster set. The values are calculated using the combined OViN dataset from 2013-2017.

Clusters	ATE						OBE						Ratio					
	Car d	Car p	Train	BTM	Bike	Walking	Car d	Car p	Train	BTM	Bike	Walking	Car d	Car p	Train	BTM	Bike	Walking
1 & 0	16.85	4.84	-3.59	-10.00	3.00	-11.10	24.24	7.06	-5.80	-11.32		-13.23	0.70	0.69	0.62	0.88		0.84
2 & 0	12.31	2.42	-3.52	-1.34	-5.29	-4.59	16.01	5.19	-5.95	-2.69	-6.73	-5.84	0.77	0.47	0.59	0.50	0.79	0.79
2 & 1	-4.11	-1.59		8.37	-8.36	5.90	-8.23	-1.86		8.63	-5.78	7.39	0.50	0.85		0.97	1.45	0.80
3 & 0	1.80		-0.93	-1.09	2.27	-2.06	5.28	1.25	-1.54	-1.80		-3.45	0.34		0.60	0.61		0.60
3 & 1	-11.94	-3.63	1.49	7.81	-2.30	8.57	-18.96	-5.80	4.26	9.52		9.78	0.63	0.63	0.35	0.82		0.88
3 & 2	-9.05	-1.87	2.31		6.17	2.71	-10.73	-3.94	4.41	0.89	6.98	2.39	0.84	0.47	0.52		0.88	1.13
4 & 0	1.57		-0.52		-0.85		2.76	1.07	-0.87	-0.28	-1.56	-1.12	0.57		0.60		0.54	
4 & 1	-13.77	-3.71	2.58	10.05	-5.90	10.75	-21.48	-5.99	4.93	11.04		12.11	0.64	0.62	0.52	0.91		0.89
4 & 2	-9.79	-1.86	2.67	1.33	3.04	4.61	-13.25	-4.12	5.08	2.41	5.16	4.72	0.74	0.45	0.53	0.55	0.59	0.98
4 & 3				1.13	-2.74	2.10	-2.52		0.67	1.52	-1.82	2.33				0.74	1.51	0.90
5 & 0	7.01	2.31	-1.48	-6.82	4.87	-5.89	13.18	2.85	-3.20	-8.70	3.76	-7.88	0.53	0.81	0.46	0.78	1.30	0.75
5 & 1	-7.78	-2.66	1.50	1.87		5.76	-11.06	-4.21	2.60	2.62	4.71	5.35	0.70	0.63	0.58	0.71		1.08
5 & 2	-2.85		1.94	-6.18	9.24		-2.84	-2.35	2.75	-6.01	10.48	-2.04	1.00		0.71	1.03	0.88	
5 & 3	3.34			-5.72	3.97	-2.47	7.90	1.59	-1.66	-6.90	3.50	-4.43	0.42			0.83	1.13	0.56
5 & 4	4.85			-7.18	7.03	-5.37	10.41	1.78	-2.33	-8.42	5.32	-6.76	0.47			0.85	1.32	0.79
6 & 0	5.48	1.34	-1.68		-2.64	-2.23	6.72	2.36	-2.57	-0.92	-2.60	-2.98	0.82	0.57	0.65		1.02	0.75
6 & 1	-9.81	-2.36	1.68	9.94	-6.57	7.13	-17.52	-4.7	3.23	10.39		10.25	0.56	0.5	0.52	0.96		0.70
6 & 2	-6.11	-1.39	1.37	1.03	1.98	3.11	-9.30	-2.83	3.38	1.77	4.12	2.86	0.66	0.49	0.41	0.58	0.48	1.09
6 & 3	2.56		-0.80	1.08	-4.02			1.11	-1.03	0.88	-2.86				0.78	1.23	1.41	
6 & 4	2.83		-0.90			-1.62	3.95	1.29	-1.7	-0.65		-1.86	0.72		0.53			0.87
6 & 5				6.52	-7.44	2.55	-6.46			7.78	-6.36	4.90				0.84	1.17	0.52

Table J.5: ATE, OBE and the ATE OBE ratio for the unweighted cluster set. The values are calculated using the combined OViN dataset from 2013-2017.

Clusters	ATE						OBE						Ratio					
	Car d	Car p	Train	BTM	Bike	Walking	Car d	Car p	Train	BTM	Bike	Walking	Car d	Car p	Train	BTM	Bike	Walking
1 & 0	7.58	2.21	-1.40	-5.21		-4.71	10.23	3.94	-2.19	-5.49		-5.26	0.74	0.56	0.64	0.95		0.90
2 & 0	5.12		-1.55	2.86	-6.01				-2.08	3.31	-4.40				0.75	0.86	1.37	
2 & 1	-3.68	-2.63		9.41	-7.43	3.78	-9.75	-2.84		8.81	-3.17	6.85	0.38	0.93		1.07	2.34	0.55
3 & 0	-4.76	-1.18		3.31		2.79	-9.19	-2.44	2.68	4.83		3.62	0.52	0.48		0.69		0.77
3 & 1	-11.63	-3.92	1.61	8.53	-1.72	7.12	-19.42	-6.38	4.87	10.32	1.74	8.88	0.60	0.61	0.33	0.83	-0.99	0.80
3 & 2	-9.84	-1.45	2.11		6.45	2.57	-9.67	-3.54	4.76	1.51	4.91	2.03	1.02	0.41	0.44		1.31	1.27
4 & 0					3.81	-5.67			1.84	-2.85	5.95	-7.55					0.64	0.75
4 & 1	-6.01	-2.08	2.52	3.51			-7.07	-4.50	4.03	2.64	7.18		0.85	0.46	0.63	1.33		
4 & 2			2.80	-5.94	11.43	-7.24			3.92	-6.17	10.36	-9.14			0.71	0.96	1.10	0.79
4 & 3	5.08	2.50		-6.38	4.72	-6.05	12.35			-7.68	5.45	-11.17	0.41			0.83	0.87	0.54
5 & 0	-8.83	-2.33	1.87	4.72		5.24	-14.72	-3.70	4.35	6.42		6.76	0.60	0.63	0.43	0.74		0.78
5 & 1	-18.02	-5.86	4.08	10.27		10.02	-24.95	-7.64	6.54	11.92	2.11	12.03	0.72	0.77	0.62	0.86		0.83
5 & 2	-13.67	-2.60	3.71	1.45	6.79	4.32	-15.2	-4.80	6.43	3.11	5.29	5.18	0.90	0.54	0.58	0.47	1.28	0.83
5 & 3	-1.83		0.94	0.91	-1.54	1.78	-5.53	-1.26	1.67	1.60		3.15	0.33		0.56	0.57		0.57
5 & 4	-9.79	-3.34		7.56	-5.17	10.5	-17.88	-3.15	2.51	9.28	-5.07	14.31	0.55	1.06		0.81	1.02	0.73
6 & 0	-6.66	-1.35	1.66	4.76	-3.84	5.42	-11.85	-2.84	3.48	6.04		5.91	0.56	0.48	0.48	0.79		0.92
6 & 1	-15.24	-3.76	3.05	10.49	-4.60	10.07	-22.08	-6.78	5.67	11.53		11.18	0.69	0.55	0.54	0.91		0.90
6 & 2	-11.31	-1.35	3.11	1.77	3.28	4.50	-12.33	-3.94	5.56	2.73	3.66	4.32	0.92	0.34	0.56	0.65	0.90	1.04
6 & 3	-1.00		0.48	0.97	-2.31	1.98	-2.66		0.80	1.21	-1.25	2.29	0.38		0.60	0.80	1.85	0.86
6 & 4	-8.43	-1.49		7.78	-8.30	10.22	-15.01	-2.29	1.64	8.89	-6.70	13.46	0.56	0.65		0.88	1.24	0.76
6 & 5	1.45		-0.35		-1.15		2.87	0.86	-0.87	-0.38	-1.63	-0.85	0.51		0.40		0.71	

Table J.6: ATE, OBE and the ATE OBE ratio for the degrees of urbanisation. The values are calculated using the combined OViN dataset from 2013-2017.

DU	ATE						OBE						Ratio					
	Car d	Car p	Train	BTM	Bike	Walking	Car d	Car p	Train	BTM	Bike	Walking	Car d	Car p	Train	BTM	Bike	Walking
2 & 1																		
3 & 1	1.67	0.79	-0.60			-1.72	3.90	1.28	-1.06		-2.15	-1.88	0.43	0.62	0.57			0.91
3 & 2			-0.50			-1.28	3.04	0.90	-1.00	-0.26	-1.21	-1.48			0.50			0.86
4 & 1	4.21	0.72	-1.15	-1.08	1.08	-3.77	8.97	2.29	-2.50	-1.99	-2.15	-4.62	0.47	0.31	0.46	0.54	-0.50	0.82
4 & 2	3.15		-1.12	-1.28	2.46	-3.48	8.12	1.91	-2.43	-2.16	-1.21	-4.23	0.39		0.46	0.59	-2.03	0.82
4 & 3	2.61		-0.67	-1.41	1.55	-2.28	5.08	1.01	-1.43	-1.90		-2.75	0.51		0.47	0.74		0.83
5 & 1	5.72	1.51	-1.95	-4.16	4.08	-5.20	12.53	3.01	-4.03	-5.37		-6.39	0.46	0.50	0.48	0.77		0.81
5 & 2	5.39	1.18	-1.86	-3.99	3.89	-4.61	11.68	2.63	-3.96	-5.55		-5.99	0.46	0.45	0.47	0.72		0.77
5 & 3	3.95		-0.96	-4.51	3.98	-3.45	8.63	1.72	-2.97	-5.29	2.40	-4.51	0.46		0.32	0.85	1.66	0.76
5 & 4				-3.02	2.85		3.56		-1.53	-3.39	2.41	-1.76				0.89	1.18	
6 & 1	12.04	3.26	-3.44	-8.20	5.15	-8.81	21.09	6.03	-5.19	-10.17		-11.52	0.57	0.54	0.66	0.81		0.76
6 & 2	11.77	2.79	-3.55	-8.33	6.81	-9.49	20.24	5.65	-5.13	-10.34		-11.13	0.58	0.49	0.69	0.81		0.85
6 & 3	10.29	2.55	-1.85	-9.06	6.37	-8.29	17.2	4.75	-4.13	-10.08	1.91	-9.64	0.60	0.54	0.45	0.90	3.34	0.86
6 & 4	5.50	2.41	-0.94	-7.07	5.26	-5.16	12.12	3.74	-2.70	-8.18	1.92	-6.90	0.45	0.64	0.35	0.86	2.74	0.75
6 & 5	5.92	1.64		-4.00	2.05	-4.95	8.56	3.02	-1.16	-4.80		-5.13	0.69	0.54		0.83		0.96

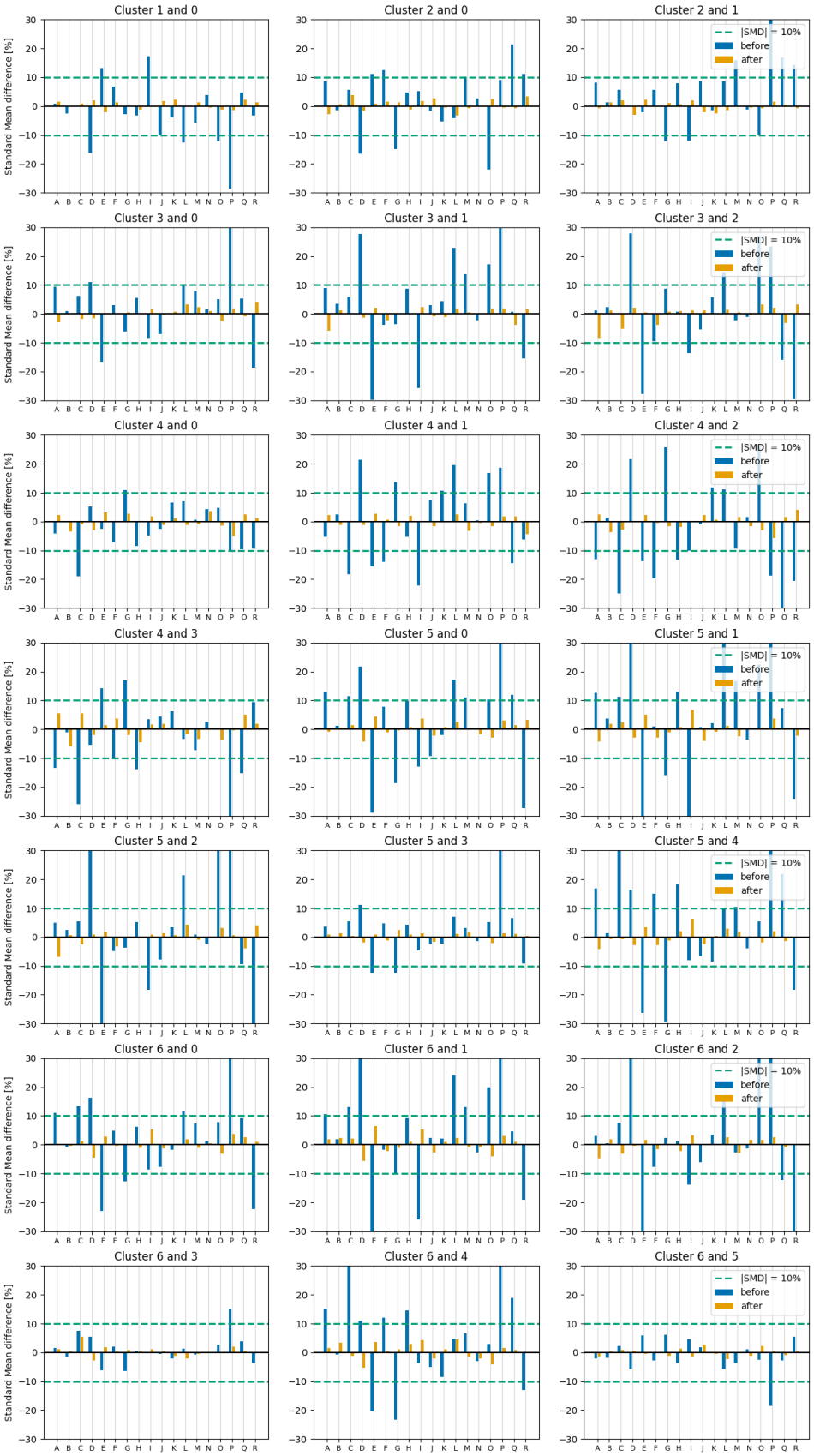

Figure J.4: SMD values before and after performing the PSM for each cluster pair and each demographic characteristic for the unweighted cluster set. The following demographic characteristics can be seen on the x-axis: A: Age; B: Gender; C: Income; D: Household size; E: 1 person household; F: 2+ person household; G: 1 parent household; H: Part time worker; I: Full time worker; J: Student; K: Primary education or less; L: lbo, vmbo; M: mbo, havo, vwo; N: other education; O: Younger than 15; P: Number of household cars; Q: Driver's licence; R: Student OV. The reference characteristics (e.g. 2 parent household for the household categories) are not included. The demographic characteristics used to calculate the SMD are based on the combined OViN dataset from 2013-2017.

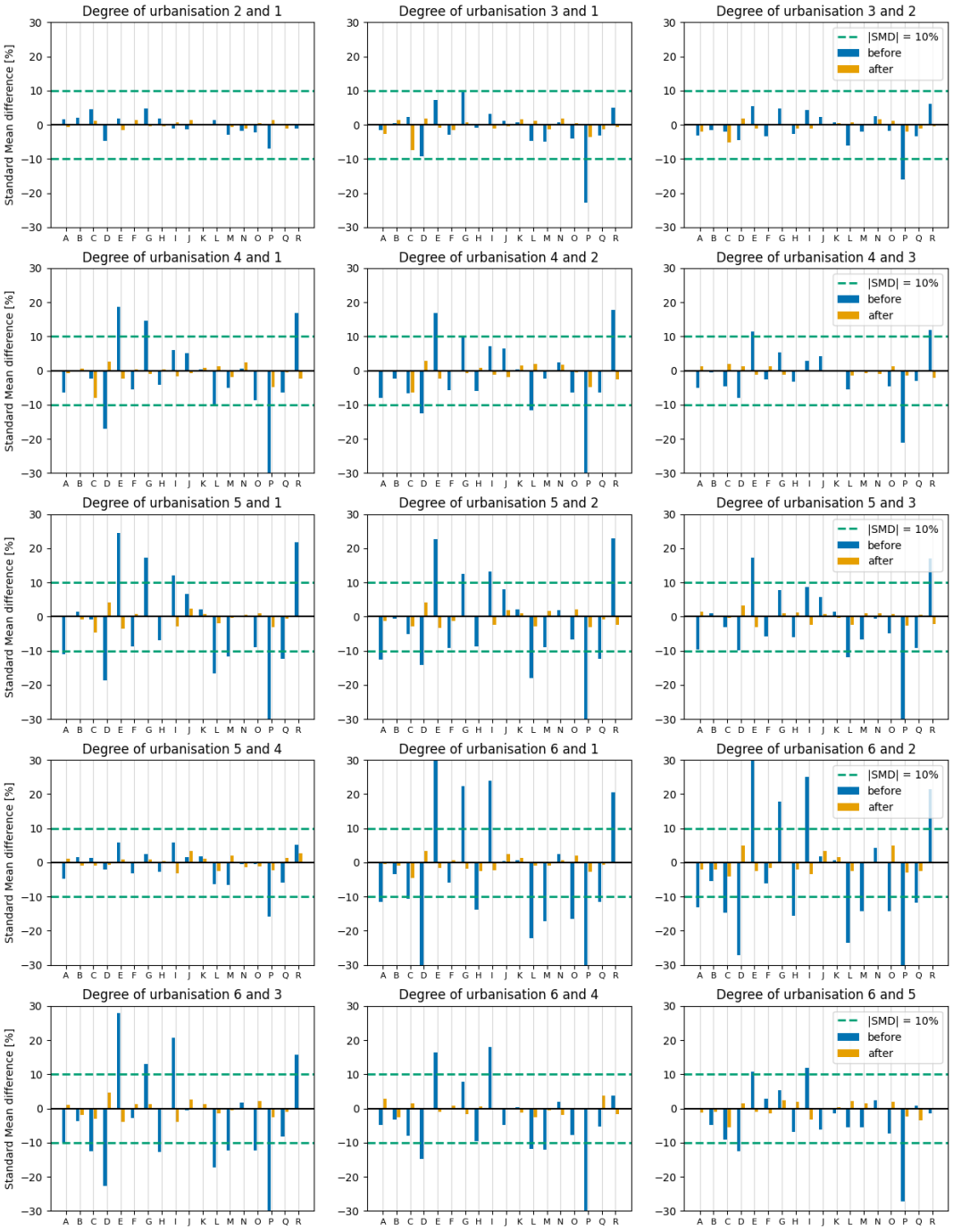

Figure J.5: SMD values before and after performing the PSM for each degree of urbanisation pair and each demographic characteristic. The following demographic characteristics can be seen on the x-axis: A: Age; B: Gender; C: Income; D: Household size; E: 1 person household; F: 2+ person household; G: 1 parent household; H: Part time worker; I: Full time worker; J: Student; K: Primary education or less; L: lbo, vmbo; M: mbo, havo, vwo; N: other education; O: Younger than 15; P: Number of household cars; Q: Driver's licence; R: Student OV. The reference characteristics (e.g. 2 parent household for the household categories) are not included. The demographic characteristics used to calculate the SMD are based on the combined OViN dataset from 2013-2017.

K

Testing propensity score matching

This appendix will give more information about tests that were done before the final PSM was performed.

First, PSM will be tested using 4 demographic characteristics with the trips as observations. After that, the zones itself and their average travel behaviour will be used as observations. This will help determining which method is the most suitable to use for the PSM.

K.1. Trips as observations

To test the PSM, it was used on a smaller scale and in a slighlty simplified form. 16 different LMS zones were handpicked. For each DU, there are 2 or 3 zones and each zone has more than 200 OViN trips to make sure the zones are representative by having enough data. The specific zones were picked semi-randomly, though interesting locations based on the exploratory data analysis were included (e.g. a zone from Leiden and a zone from Zoetermeer).

The 16 zones were divided into 3 clusters using hierarchical clustering, based on 4 different spatial environment characteristics (distance to city centre, road density, entropy, distance to train station), see figure K.1 for the corresponding dendrogram and figure K.2 for the clusters. All the variables were scaled to a range from 0 to 1. Interestingly, the zone in Zoetermeer and Rotterdam were placed in the same cluster (cluster 2), while the zone from Leiden was placed in the same cluster a zone from Utrecht, Amsterdam and The Hague (cluster 1).

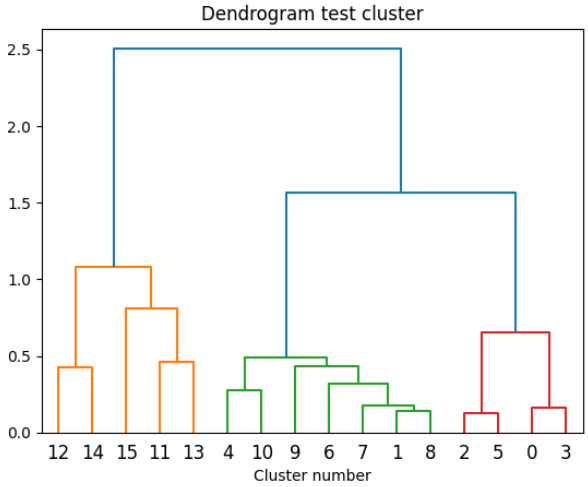

Figure K.1: Dendrogram hierarchical clustering for the test clusters. The sources for the variables used in the clustering can be found in table 4.1.

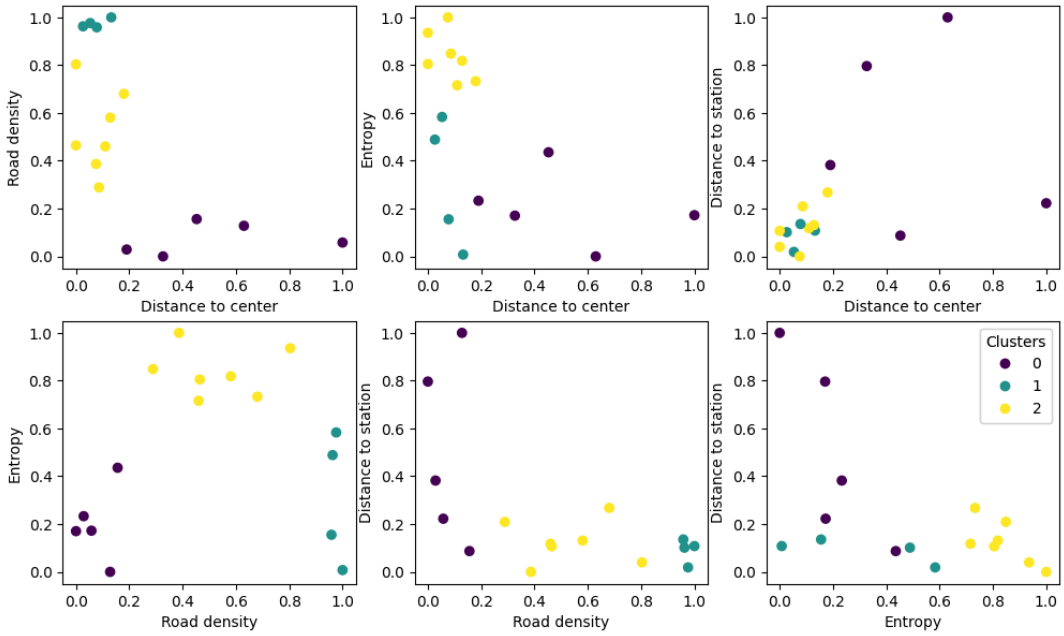

Figure K.2: Scatter plot of the different variables used in clustering. All variables are scaled between 0 and 1. The sources for the variables used in the clustering can be found in table 4.1.

The percentage of car use for trips departing from a selected zone was used as indicator for travel behaviour. When comparing the different distributions between the clusters using a t-test, all p-values were very close to 0 and way below 0.05.

After that, the propensity score was calculated using logistic regression on 4 demographic characteristics (age, income, number of cars, household size). The distribution of the propensity scores for each cluster pair is plotted in figure K.3. As shown in the figure, there is sufficient overlap in propensity scores between the different clusters, so the matching has a high chance of succeeding.

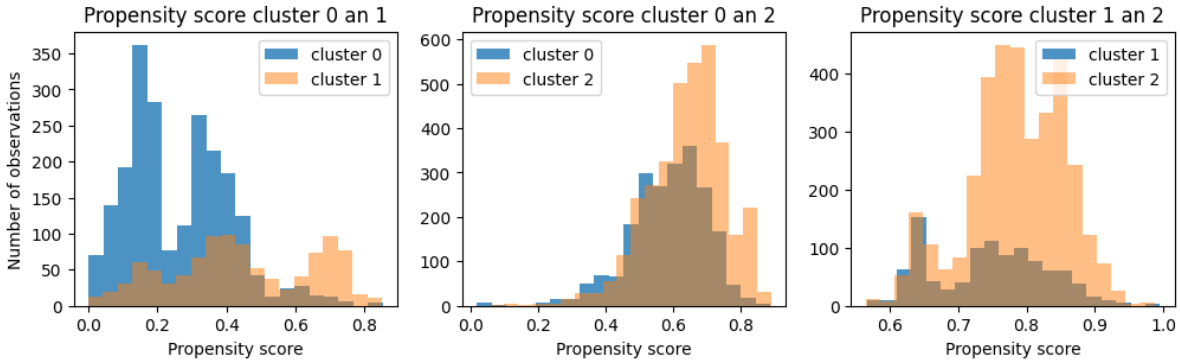

Figure K.3: Distribution of propensity scores of different cluster pairs. The demographic characteristics used to calculate the propensity scores are based on the combined OViN dataset from 2013-2017.

The different clusters were matched using a caliper (maximum difference between propensity scores) of 0.01. For all three matched (cluster 0 and 1, cluster 0 and 2, cluster 1 and 2) 70-90 % of the data points of the smallest clusters were kept. This gave cluster sizes of 711, 1829 and 964 data points of the three cluster pairs.

The t-test was performed again on the percentage of car use. This time both cluster 0 and 1 and cluster 1 and 2 were different (p-value very close to 0, although larger than the previous time). Cluster 0 and 2 however, now had a p-value of 0.275, which is larger than 0.05. Thus it cannot be assumed that the car use between those two clusters is different based on this data.

Figure K.4 shows the differences in modal split between the different clusters, corrected by demography. Figure K.5 shows the modal split for the different clusters, uncorrected by demography for both OViN and LMS. It shows that cluster 0 and 2 have fairly similar mode use for all modes, except public transport according to the corrected OViN data. The differences between those two clusters are a lot larger for the uncorrected OViN and LMS data, which implies that the demography has a large effect on the differences in travel behaviour.

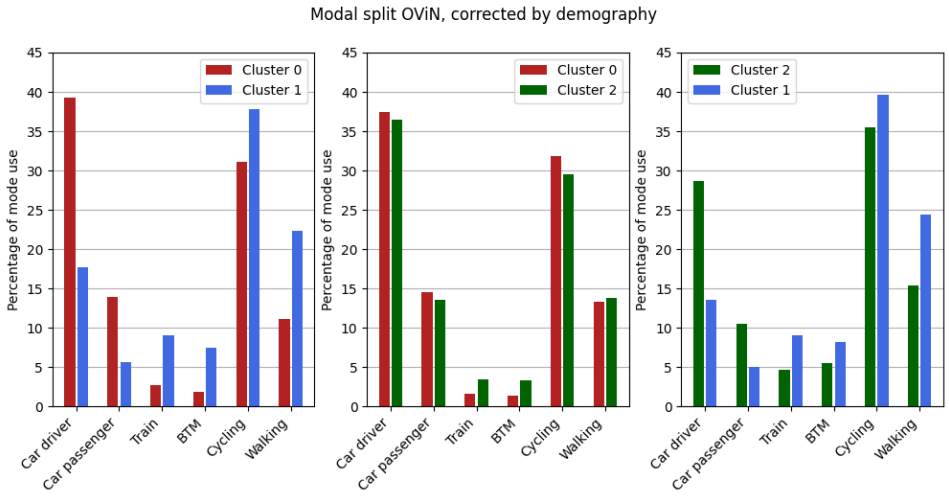

Figure K.4: Travel behaviour of different cluster pairs according to the combined OViN dataset from 2013-2017, after correcting for demography.

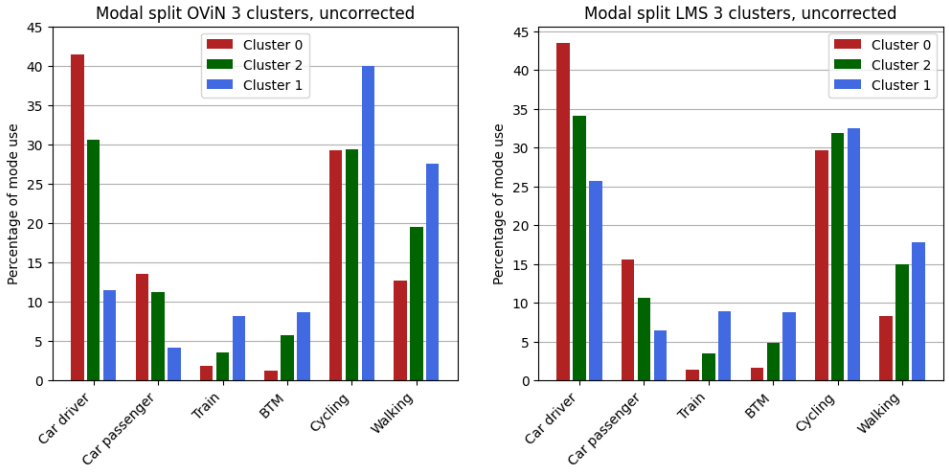

Figure K.5: Travel behaviour OViN and LMS for different clusters without correcting for demography. The LMS data is based on RWS WVL (2018c) and the OViN data on the combined OViN dataset from 2013-2017.

Cluster 1 also shows some interesting results. According to uncorrected OViN the car drivers make up of around 12 % of the trips, which is closer to 17 % after correction, when comparing with cluster 0 and around 14% when comparing with cluster 2. According to LMS however, it accounts for more than 25 % of the trips, which is even higher than the corrected OViN results. Cluster 1 still shows very large differences in travel behaviour with the other 2 clusters after correction, which implies that the spatial environment plays a large roll in the differences in travel behaviour, which could be a reason why the LMS, for example, largely overestimates car use for that cluster.

After matching, the SMD was lower than 10% in all cases, except for the household size for the clusters 0 and 1, which is just above the 10%. See figure K.6.

This test only included 16 clusters, instead of 1406, and used no weight factors for the observations.

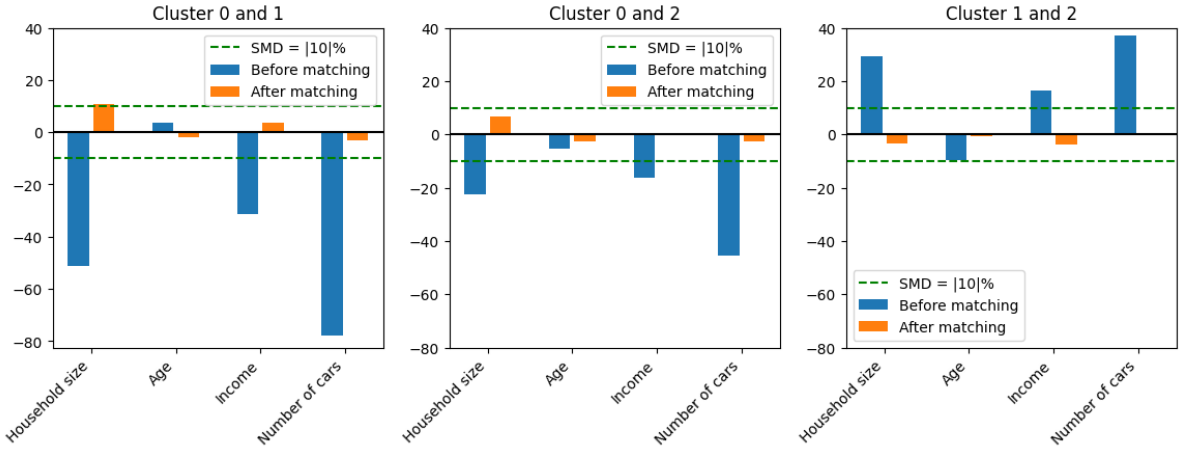

Figure K.6: Standard mean difference for the different cluster pairs. The demographic characteristics used to calculate the SMD are based on the combined OViN dataset from 2013-2017.

Table K.1: ATE, OBE and ATE to OBE ratios for the test clusters. The modal splits are based on the combined OViN dataset from 2013-2017.

mode	Cluster 0 & 1			Cluster 0 & 2			Cluster 1 & 2		
	OBE	ATE	Ratio	OBE	ATE	Ratio	OBE	ATE	Ratio
Car driver	-29.97	-21.50	0.72	-10.91	-0.99	0.09	19.06	15.10	0.79
Car passenger	-9.38	-8.27	0.88	-2.33	-0.99	0.42	7.05	5.37	0.76
Train	6.35	6.33	1.00	1.78	1.78	1.00	-4.57	-4.46	0.98
BTM	7.44	5.61	0.75	4.54	1.89	0.42	-2.90	-2.76	0.95
Bike	10.72	6.64	0.62	0.13	-2.20	-16.75	-10.59	-4.23	0.40
Walking	14.85	11.19	0.75	6.80	0.52	0.08	-8.05	-9.03	1.12

Presumably, with a sample size that is almost 100 times larger, there will be enough observations to find the right matches. Of course, for the real PSM, all demographic characteristics will be included instead of only 4.

Table K.1 shows the ATE, OBE and ATE-OBE ratio for the different clusters. The ratio is a measure for how much of the differences in travel behaviour that are observed come because of the spatial environment. When comparing cluster 0 and 1, all values are above 60%, which means that a large part of the differences come from the spatial environment. The ratio between clusters 1 and 2 for walking shows a value higher than 100%. This implies that the effect of the spatial environment might even be underestimated when the demography is not taken into account. The p-values of these ATE and OBE values have not been calculated, which means that not all ratios are accurate (e.g. bike ratio for cluster 0 and 2).

All in all, the first results from these test seem positive and imply that indeed the spatial environment plays a large roll in travel behaviour and that this method will be suitable to determine the effect of the spatial environment for the final clusters.

K.2. Zones as observations

It was also explored if it is possible to use the average travel behaviour of zones as observations for the PSM. This way, it would be possible to compare OViN and LMS before and after matching.

Using hierarchical clustering, 3 clusters were made based on the same 4 D-variables (distance to city centre, road density, entropy, distance to train station). After that, the propensity scores were calculated using logistic regression using the same 4 demographic characteristics (age, income, number of cars, household size).

When comparing the propensity scores of the different clusters, there was already less overlap compared to the previous test that used the observations of only 16 clusters. See figure K.7. Especially

the clusters 1 and 2 barely have any overlapping observations.

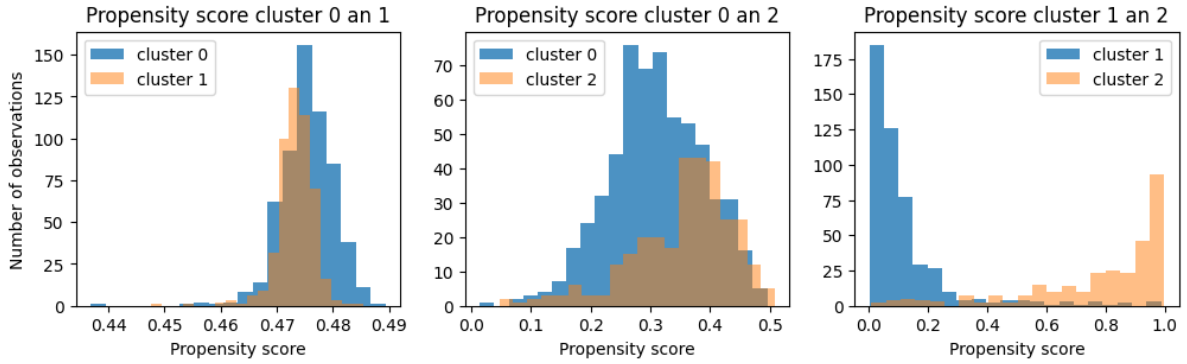

Figure K.7: Distribution of propensity scores of different cluster pairs when using the zones as observations. The demographic characteristics used to calculate the propensity scores are based on RWS WVL (2020) and CBS & ESRI Nederland (2019).

After that, the PSM was done and the SMD was calculated, see figure K.8. This figure shows that after the matching, the SMDs for most demographic characteristics were still a lot higher than 10 %. Presumably, when using more clusters than 3 and more demographic characteristics, it would be even more difficult to perform the matching. From this test, it can be concluded that using the zones as observations is not suitable and the trips will be used.

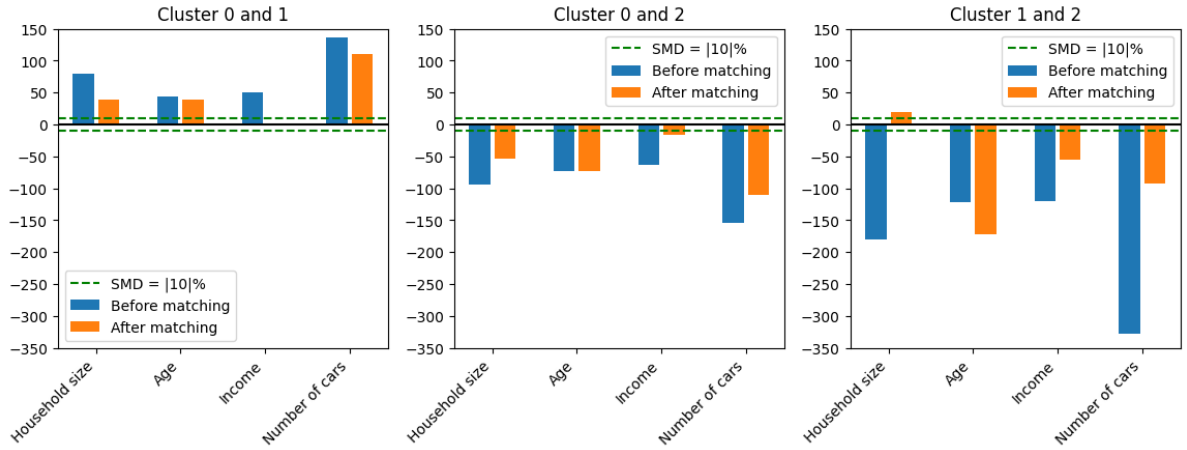

Figure K.8: The SMD for all cluster pairs when using the zones as observations. In almost all cases, the SMD is larger than 10% after matching. The demographic characteristics used to calculate the SMD are based on RWS WVL (2020) and CBS & ESRI Nederland (2019).

Data management plan and HREC checklist

At the start of the thesis a data management plan was maded and a Human Research Ethic checklist was filled in. This was needed because the OViN data contains personal data (e.g. postal code, income, age, gender) and approval by the Humans Research Ethichs Committee (HREC) was needed before the data analysis could be started. The approval was given and a screenshot of the approval is given in figure L.1.

Date 19-Mar-2024
Contact person Grace van Arkel, Policy Advisor
Academic Integrity

TU Delft Delft University of Technology

Human Research Ethics
Committee TU Delft
(http://hrec.tudelft.nl)

Visiting address
Jaffalaan 5 (building 31)
2628 BX Delft

Postal address
P.O. Box 5015 2600 GA Delft
The Netherlands

Ethics Approval Application: The effect of the degree of urbanisation on travel behaviour
Applicant: Timmerman, Janine

Dear Janine Timmerman,

It is a pleasure to inform you that your application mentioned above has been approved.

Thanks very much for your submission to the HREC which has been approved.

In addition to any specific conditions or notes, the HREC provides the following standard advice to all applicants:
• In light of recent tax changes, we advise that you confirm any proposed remuneration of research subjects with your faculty contract manager before going ahead.
• Please make sure when you carry out your research that you confirm contemporary covid protocols with your faculty HSE advisor, and that ongoing covid risks and precautions are flagged in the informed consent - with particular attention to this where there are physically vulnerable (eg: elderly or with underlying conditions) participants involved.
• Our default advice is not to publish transcripts or transcript summaries, but to retain these privately for specific purposes/checking; and if they are to be made public then only if fully anonymised and the transcript/summary itself approved by participants for specific purpose.
• Where there are collaborating (including funding) partners, appropriate formal agreements including clarity on responsibilities, including data ownership, responsibilities and access, should be in place and that relevant aspects of such agreements (such as access to raw or other data) are clear in the Informed Consent.

Good luck with your research!

Sincerely,

Figure L.1: Approval HREC