Predicting bus ridership in the Netherlands

Master thesis submitted to Delft University of Technology

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in **Engineering and Policy Analysis**

Faculty of Technology, Policy and Management

by

Sam de Boer

Student number: 4391128

To be defended in public on September 17th 2021

Graduation committee

Chairperson	: Dr.ir. M. Kroesen, Section Engineering Systems and Services
First Supervisor	: Dr.ir. M. Kroesen, Section Engineering Systems and Services
Second Supervisor	: Dr.ir. C. van Daalen, Section Policy Analysis
External Supervisor	: Ir. W.J. te Morsche, Movares Nederland

Executive summary

The Netherlands suffers from a traffic congestion problem, which is a cause of both economic damages and high NOx and CO2 emissions. Today, the bus accounts for only a small percentage of all trips made in the Netherlands (less than 4%) but the bus plays an important role as an access and egress mode for train travellers. Short access and egress times are considered as crucial criteria for choosing to travel by train. Thus without a proper bus network, travelling by train will become less attractive. Furthermore the Dutch population is both growing and more mobile than ever, but space to construct new roads is limited. These developments mean that the current transport system in the Netherlands will not be sustainable in the future. To prepare it for the future sustainable public transport will be essential.

However it is complicated to design a solid bus network as the public transport sector depends heavily on subsidies. For bus operators to allocate their resources efficiently, good information is key. Direct Ridership Models (DRMs) are rising in popularity as they can provide useful information to bus operators. DRMs are statistical models that are able to capture the relationship between surroundings (such as car ownership in the surrounding neighbourhood or the presence of a university) and travel demand. This way first cut predictions can be made for new bus stop locations. Furthermore, the expected number of travellers can be predicted for existing bus stops when there is a change in the surroundings of the bus stop, which allows bus operators to adjust their services accordingly.

In this study, DRMs are used to explain and predict bus ridership for the concession area Arnhem-Nijmegen. The independent variables that are included in the models can be categorised into three groups: demographics, built environment and level of service. Three DRMs are developed based on different regression models. One model is the commonly used OLS. The other two models, Spatial Lag X model (SLX) and Spatial Error Model (SEM), are able to take spatial relationships between bus stops into account.

The spatial models showed an improvement over the OLS model in terms of explanatory power. With 34 independent variables, the OLS model could explain 71.4% of the bus ridership. The SEM was able to explain 72.4% and the SLX 73.4%. But more importantly, some of the independent variables that were found significant related to ridership in the OLS model turned out to be not significant in the SEM model, of which car ownership was the most surprising one. For Arnhem-Nijmegen, it could be further looked into if the transportation by bus to parks, sport facilities, amusement parks and hospitals could be facilitated more to increase the bus usage as these variables are found to influence bus ridership significantly. Furthermore, it should be examined why P+R's are not found to affect bus ridership significantly.

The prediction accuracy of the OLS model and the SLX model was also studied. The predictions of the SLX model turned out to be more accurate, but in general the prediction results were not satisfying. For more than half of the bus stops in Arnhem-Nijmegen, the predicted value of the SLX model was 50% more or 50% less than the actual value. The reason for this is the existence of non-stationary relationships, which means that the relations between ridership and the independent variables can vary based on the locations of the stops. This study contained bus stops in urban areas and in rural areas and it can be expected that for example the influence of a hospital in a village on bus ridership is smaller than the influence of a hospital in a city.

In this study, the existence of non-stationary relationships between regions is demonstrated while studying if a DRM based on Arnhem-Nijmegen can be generalised to the region Groningen-Drenthe. It turned out residential land-use has more influence on bus ridership in Groningen-Drenthe while hospitals and universities have a higher impact on ridership in Arnhem-Nijmegen. The non-stationary relationships make it not advisable to generalise the DRM based on Arnhem-Nijmegen to different regions.

In scientific literature, nothing has been written about the practical applications of DRMs for policy makers. As it was assumed DRMs can be useful for policy makers, three policy makers were interviewed for their insights on the applications. One of the application mentioned is preventing unsafe traffic situations, as DRMs allow to update ridership predictions when circumstances around the bus stop change. For example, when a bus stops is designed for a low number of travellers and a new neighbourhood is developed close to it, the increase in travellers could result in unsafe

situations. With a DRM the increase in travellers can be predicted, which makes it possible to redesign the bus stop. Other applications that were identified in the interviews are: examining if a bus stop is performing below its potential and substantiating plans for new routes, new stop locations and relocating bus stops. Overall, the benefit of DRM is widely acknowledged and in combination with personal expertise, it could lead to optimised decision making.

In the end, it is demonstrated how a DRM can be used to identify bus stops which do not fulfil their potential and suggestions are provided how the number of rides can be increased. The stops Europalaan in Renkum, Kerkstraat in Weurt and Beemdstraat, Fransestraat & Hatertseweg in Nijmegen turned out be under performing and several improvement suggestions were made.

Although the prediction accuracy of the models were rather poor, future studies can use this study as a basis for models based on a more homogeneous study area, such as only bus stops in cities. Sub-setting the model is expected to have a major positive influence on the prediction accuracy. It would be interesting to see if a DRM based on a city can be generalised to other cities as well. Furthermore, the model can be complemented with more independent variables. Variables that encourage car use over bus use, such as the difference in travel time to the city center or the distance to the closest highway ramp, could be valuable additions.

Preface

In the last two years studying EPA I have discovered my love for programming in Python and Rstudio and during an elective I came to the conclusion that I would like to work in the mobility section. Therefore, I was (and still am) very glad that I got offered this assignment by Movares. Although I wonder if I ever can look at a bus stop without thinking what the number of rides will be.

I would like to thank Maarten Kroesen for his expertise on statistics and for his efforts trying to let me focus on my report instead of the model. I would like to thank Els van Daalen, who gave a lot of helpful feedback for my story line and I would like to thank Wietse te Morsche, for his extensively guidance and teaching me the ins and outs of the mobility world. Furthermore, I would like to thank OV Oost and OV Bureau Groningen Drenthe who provided me the smart card data, something that I understand is quite unique.

Sam de Boer

Rotterdam, September 2021

Contents

List of Figures

List of Tables

1 Introduction

1.1 Problem statement

A solid transportation infrastructure is an important driver of prosperity for a country. A good infrastructure network gives citizens better access to healthcare and education, more high-quality employment opportunities and guarantees reliable logistical operations for businesses. (Puentes, 2016). However, these days road congestion is more often the rule than the exception in the Netherlands. And congestion comes with a price, KIM (2019) has estimated the total costs of highway network congestion to be between €3.3 billion and €4.3 billion in 2018. Included in these costs are costs that are directly measurable, like extra travel time and the unreliability of the travel time, and costs which are harder to measure, such as lower productivity caused by arriving late at work. Besides economic damages, the environment is damaged as well. Den Boer and Vermeulen (2004) showed that congestion causes high NOx and CO2 emissions. The NOx emission of a truck in a traffic jam for example doubles compared to a truck on a road with light traffic . This is a cause for concern as NOx is harmful to humans, while CO2 emissions contribute to global warming.

Highway traffic jams take up one third of total congestion in the Netherlands, and 50% of the congestion take place in the cities (Kuiken, 2020). In large cities, drivers therefore have to take a delay of more than 20% into account when planning their trips (Folkert, 2019). Furthermore emission levels from congested traffic in cities are especially harmful to public health, due to the higher public density (as highlighted by Kuiken).

These are just some of the many reasons why the transportation system in the Netherlands needs to become more sustainable. To achieve this, public transport is essential (Costa, Fontes, Costa, & Dias, 2015).

1.1.1 Dutch travel behaviour

One of the things the Netherlands is known for is its bike-friendly culture. In 2018 there were 17 million people and 23 million bicycles in the Netherlands, meaning there are 1.35 bikes per person in the country (KIM, 2018). But the car is still the most used mode of transport in the number of trips (29% of all trips), the bike is a close second (26%) and walking (23%) is in the third place (KIM, 2019). When it comes to public transport the train accounts for 13% of the trips made and the tram, bus and the metro combined only for 4%.

But in contradiction to the car, the kilometers travelled per person per year are rising for the train, bus, metro and tram. In the period from 2010 till 2017 the number of kilometers travelled by car declined by 5.5% (KIM, 2019). The average number of kilometers travelled by train per person per year rose from 1,030 in 2010 to 1,132 km in 2018, an increase of 9.9%. And in an even shorter period of time (from 2014-2018) the average number of kilometers travelled by tram, metro and bus rose with 3.5%. The decline in car usage and the increase in usage of public transport is in line with policies being pushed by cities. More and more cities are trying to reduce the number of cars driving in the city boundaries, with Utrecht and Amsterdam just being some examples of the many. (Gemeente Utrecht, 2020) (Gemeente Amsterdam, 2013).It used to be normal in urban planning to plan between 1 and 1.8 parking spaces per residence (Crow, 2018), but a recently built neighbourhood in Utrecht is completely car-free. Parking garages have been banished to the outskirts of the neighbourhood, and only offer one parking space for three residences (Bremmer, 2020). Meanwhile the city of Amsterdam is planning to remove more than 10,000 parking spaces inside the city ring before 2025 (Koops, 2019). And cities are not just discouraging car use. Cities such as Nijmegen and Arnhem have made improving public transport infrastructure a number one priority in order to improve traffic flows (Gemeente Nijmegen, 2018) (Gemeente Arnhem, 2016).

There is a clear correlation between the reason people are traveling, and the mode of transport they use for their trip. The main transport mode for commuting is the train, while for travelling to schools and universities the bus,metro and tram are the most used modes of transport (CPB and KIM, 2009). School children and students make up 15% of all travellers in the Netherlands,

but they make up one third of all kilometers travelled by train and almost half of the kilometres travelled by bus, metro and tram. The elderly on the other hand are travelling less by train. The percentage of trips made by the elderly by bus, tram or metro is similar to the percentage of an average Dutch person (CPB and KIM, 2009). There is also a clear division between household income and public transport usage. Households with an income below €22.500 per year, good for one sixth of the number of households, travel a quarter of all public transport kilometers.

1.1.2 The increasing importance of the bus

Besides decreasing congestion and thus improving accessibility, a well-functioning public transport system comes with additional advantages, such as fueling business growth, keeping cities green and livable, creating more efficient cities and driving more equal opportunities for society (van Oort, van der Bijl, & Verhoof, 2017). All these advantages have been cited as targets over the years that the Dutch government should strive for (CPB and KIM, 2009).

Although the role of the bus may seem small at first sight in the Netherlands, the transport mode is an important enabler for the train. When the main transport mode of travellers is the train, 19% of them take the bus, metro or the tram to the train station. After the train trip, 33% of the travellers continue by bus, metro or tram. So without a proper bus, tram or metro network in a city, the train will become much less attractive. This is because frequencies and access & egress times are considered the most important quality aspects of public transport (KiM, 2018). With a growing population and thus an increase in the number of travel movements and at the same time a lack of space to build more roads, one can anticipate that the role of trains and other public transport will grow in importance in the future.

1.1.3 Dutch public transport system and DRM

The Netherlands is divided into 34 concession areas for public transport (Crow, 2020). Almost every area sets out a tender and the public transit company with the best level-of-service (LoS) and the lowest price is then granted the exclusive right to provide the public transport services for a couple of years (Rijksoverheid, n.d.). This indicates one of the key factors in public transport: efficiency. Efficiency matters because public transport is heavily subsidized to keep public transport accessible for travellers (Niaounakis, Blank, & Veeneman, 2016), and citizens do care how their tax money is being spent. Public transport companies which participate in a tender have to hand in their service plans for the concession area to transport authority. The companies should address in their plans how they are going to allocate their resources and the plans should be as accurate and efficient as possible (Chu, Polzin, Pendyala, Siddiqui, & Ubaka, 2006).

To achieve this, information is key. A model which can generate useful information and which has gained popularity in professional circles in recent years is the Direct Ridership Model (DRM). DRMs make it possible to capture the relationship between the surroundings of a bus stop (such as car ownership or the presence of a hospital) and travel demand (Dill, Schlossberg, Ma, & Meyer, 2013). Some studies also showed that DRMs can also be used for quick first-cut predictions of the number of travellers for (potential new) bus stops (Upchurch & Kuby, 2014). When companies are able to predict bus ridership accurately the quality of their tender plans improve and they have better chances of winning the concession. A DRM can help with this. It helps transport companies to adjust their services according to changes in the neighbourhood. For instance, in the Netherlands more and more offices are transformed into residences (Bouma & Bontjes, 2019). Knowing this and knowing which kind of people are going to live there makes it possible to predict the expected number of travellers, when a DRM is applied. Which allows for the public transport companies to design a network that attracts as many bus travellers as possible.

1.2 Research problem

1.2.1 Research gap

Although many studies stress the importance of quantitative predictions of bus ridership at the stop level as a important requirement for sustainable city development (for instance Chu (2004), Cervero et al. (2010) and Guo and Huang (2020)), most of the studies which apply a DRM only use it for explaining and examining what the relationships are between surrounding variables and bus ridership at stop level. The prediction accuracy of DRMs is rarely tested (Taylor & Fink, 2013). The studies that do often have a relatively small sample size, smaller than the ones seen in DRM studies which are only used for explaining.

But DRMs do show promising results, for instance Chow, Zhao, Liu, Li, and Ubaka (2006) and Dill et al. (2013) succeeded in explaining around 70% of the bus ridership at the stop level. Hensher (2008) however warns that these results cannot be extrapolated to different countries, as the relationships between the surrounding variables and bus ridership can vary between different locations. In the Netherlands this applies even more due to the extraordinary high bike usage.

There has also been Netherlands-focused research conducted by Kerkman, Martens, and Meurs (2015). They succeeded in explaining 77.7% of the bus ridership in the Dutch Arnhem-Nijmegen region. But the results were not translated into a predictive model and the possibility of generalization of these results to the rest of the Netherlands is not studied. In fact, generalisation is seldom discussed in the scientific literature. Furthermore, the study of Kerkman et al. can possibly be strengthened by taking spatial relationships into account, which means that areas close to each other show similar behaviour (a more in-depth explanation is given in section 2.3.2) For the studies that do examine prediction accuracy, Guo and Huang (2020) takes spatial relationships into account but unfortunately it does not specify the difference in prediction accuracy between models with spatial relationships and models without.

Apart from the lack of information on prediction accuracy of DRMs, especially for large study areas and the possibility of generalizing results to other regions isn't discussed, there is another knowledge gap. We have seen DRMs can help public transport companies and the author of this study believes that DRMs can be valuable to policy makers as well. But existing studies do not provide insights on applications of DRMs for policymakers.

1.2.2 Research objective

This study has two objectives: one quantitative and one which is qualitative in nature. The first objective is to discover if it is possible to develop a model for the Netherlands, based on a selected study area, which is able to make a first cut prediction on the number of bus ridership at stop level. The qualitative objective is to develop more insights in travel behaviour by looking at the relationships found in the model and to see if the prediction model could prove useful to policy makers. These objectives are translated into the following research question:

"What is the influence of variables on bus ridership at stop level and what is the predictive accuracy and the usability of the model for policymakers?"

Grosvernor (2000) warns that before a quantitative research can be conducted, a basic understanding of travel behaviour is required so that the researchers know what to examine. That is why the first sub-question is about identifying which variables are known based on scientific literature to influence bus ridership. Furthermore, the first sub-question is about the methods that are used to explain and predict bus ridership. The second question examines in a statistical manner the actual influence of the variables of sub-question one on the bus ridership at stop levels for the area Arnhem-Nijmegen. The third sub-question is to discover if the model can be used for predictions within Arnhem-Nijmegen. DRMs cannot be generalised to other countries and because it is for example likely that low car ownership affects bus ridership more in a village than in a city (where most of the facilities are nearby). Therefore, the fourth sub-question is about finding out if a model, based on Arnhem-Nijmegen, can be used in different parts of the Netherlands.

Because the value of DRMs for transport companies is known and it is assumed DRMs can be useful for policy makers as well, the last question is about finding the applications of DRMs for policy makers and about the insights the developed model can provide.

1. Which variables and methods are found in literature to explain and predict bus ridership?
2. Which variables influence bus ridership in the concession area Arnhem-Nijmegen?
3. What is the prediction accuracy of the model?
4. Can the results of a prediction model, based on Arnhem-Nijmegen, be generalised to other regions in the Netherlands?
5. What are the applications of ridership prediction models for policy makers and which insights can the model provide for Arnhem-Nijmegen?

According to Taylor and Fink (2013) testing for the predictive ability of transit ridership models is often neglected, with most of the researchers only examining the explanatory power of their model. This research will have a close look to the predictive power of the model and will test if it is possible to develop a model for the Netherlands based on the concession area Arnhem-Nijmegen. This kind of extrapolating is rarely tested. In comparison to Kerkman et al. more variables will be tested for their explanatory potential for bus ridership. Special attention will be paid to what destinations a traveller can reach within a certain amount of time from a bus stop. In comparison with other predictive DRM studies, this study will have a significantly larger sample size and will estimate two models: the often seen Ordinary-Least-Squares test (OLS) and a spatial lag x model. The latter is able to deal with spatial auto-correlation between ridership and the neighbouring independent variables. The prediction accuracy of the two models will be discussed to see which model performs better and if including spatial auto-correlation is a valuable addition to prediction models for bus ridership. Afterwards, three policymakers will be interviewed in order to record their views on the applications of a DRM.

1.2.3 Research approach

To answer the first sub-question, this research will start with a literature review where existing ridership models will be discussed. Explicit focus will be on which variables are used to explain transit ridership and which methods are used and why. Afterwards, inferential statistics will be applied to answer the second, third and fourth research sub-questions. Inferential statistics are often used to examine social phenomena and applying them makes it possible to draw conclusions for the whole population based on a sample group. Besides a quantitative approach, a qualitative approach can also be used to examine social phenomena. This approach gives a deeper understanding of the complexities of travel behaviour and it is possible to identify specific (regional) explanatory variables for bus ridership, which would not be possible with quantitative research (Beirão & Cabral, 2007). Common means to retrieve this information are interviews, focus groups and field research observations. The downside however is that these means are all very time consuming and expensive (Clifton & Handy, 2001), especially when considering for example that interviews should be conducted nationwide to be able to predict the local bus ridership at stop level.

A quantitative approach comes with downsides too. Only analysing data does not tell the story behind the data, there could be important contextual details missing in the data (Clifton & Handy, 2001), such as weather conditions or social events in the area. To deal with this limitation, the time window of the data analysed should be long enough to make the model less susceptible to these contextual factors. After the relationships between various variables and bus ridership have been described, the prediction accuracy of the model is evaluated to answer the second sub-question. This is done by comparing the predicted values with the observed values in Arnhem-Nijmegen. To examine if the results of the prediction model, based on Arnhem-Nijmegen, can be generalised to other regions in the Netherlands the strength of the relationships between

the explanatory variables and bus ridership in Arnhem-Nijmegen is compared to the strength of the relationships between the same explanatory variables and bus ridership in the concession area Groningen-Drenthe.

A qualitative approach is employed to answer the first part of the last sub-question. In scientific literature the practical applications of a DRM are described very briefly. To retrieve more insights in these practical application three Dutch policy makers will be interviewed for their views on the applications of a DRM. To answer the last part of the sub-question, the results of the second sub-question are examined to retrieve insights and one application identified during the interviews will be worked out in detail.

2 Literature review of bus ridership models

There have been various studies explaining and trying to predict transit ridership. The focus of this chapter will be on these studies, which examine ridership at stop level for busses, while studies of other transit modes can contain interesting insights too. Search terms that were used were "predicting patronage at stop level", "public transit ridership model" and "estimate bus ridership". Other search terms contain a combination of a variable that is expected to affect bus ridership and "transit ridership", such as "shareable e-scooters and transit ridership". Articles on predicting ridership based on live data such as mobile phones were considered to be out of the scope for this research. The articles that are described in this chapter are mainly retrieved from the data bases Taylor & Francis, ResearchGate, Springer, Wiley and Sage. Special attention was paid to the models studies are using and which variables the researchers included in their studies. A quick scan through the literature learns that there are mainly four categories of explanatory variables for transit ridership: demographics, built environment, level-of-service and prices. These categories will be separately discussed. Afterwards, the models that can be used to explain or predict bus ridership will be discussed.

2.1 Variables

Below, the variables that are described in literature as likely to influence bus ridership are discussed. The relationship between variables and bus ridership are presented in tables for the demographics, built environment and level of service categories, as variables in these studies appear often in multiple studies. In the tables, a distinction has been made between studies which standardize their results and the ones that do not. For the first ones, it is possible to say something about the relative explaining power of the variables in comparison with the other variables in the same study. For these studies, the top five variables in explanatory power (so in absolute values) have been assigned double pluses or double minus signs. Furthermore, when an explanatory variable had a significance level above 0.1, the relationship between the variable and ridership is considered as not significant. See section 3.4.2 for an explanation of the significance level.

2.1.1 Demographics

One of the first studies on transit ridership used demographics as set of explanatory variables. Explanatory variables, also named independent variables, are the variables that are used to explain bus ridership. The study included a causal model and was conducted by Dajani and Sullivan (1976). Demographics have since then been a recurrent category in later studies. Looking at figure 1, it is striking that in many studies a negative relationship between *income* and *ridership* was found. This means that the higher the income of a household, the lower the transit usage (Kerkman, Martens, & Meurs, 2018) (Pulugurtha & Agurla, 2012) (Chu, 2004) (Chu et al., 2006) (G. Thompson, Brown, & Bhattacharya, 2012) (Zhao, Chow, Li, & Liu, 2005). A similar relationship is the positive relationship between the percentage of *households with annual income below poverty* and *ridership* (Dill et al., 2013). For the variables *households without a vehicle* and the percentage of *female residents* several studies disagreed. Dill et al. (2013) were the only ones that found the counter-intuitive positive relationship between the *percentage of households with cars* and *ridership*. They named various reasons to explain this relationship, such as *households without a car*, that do not live below the poverty line , are not exceptionally high transit users or zero-vehicles households could be located in city centres where destinations are more within walking or biking distance. In general, counter-intuitive signs could be an indication of multicollinearity (Zhao et al., 2005). For the influence of the percentage of *women residents*, Chu did find contradictory results in his study at the West Coast of the US (2006) and the East Coast (2004). This could be caused by the sample sizes, as it is likely the percentage of *women* will be close to 50% most of the times and when a sample size is large enough relationships between variables are always likely surface. (MSU, n.d.).

Table 1: Demographics

	Dill et al. (2013)	Kerkman, Martens & Meurs (2015)	Pulugurtha & Agurla (2012)	Ryan & Frank (2009)	Chu (2004)	Chu (2007)	Hunt et al (1986)	Thompson (2012)	Zhao et al. (2005)
Std	No	Yes	No	Yes	No	No	No	No	Yes
Residents	+					+	+	+	
Female (%)				+	+	-			
Hispanic (%)				ns	+			ns	-
Asian (%)			+						
White (%)	-			–	-		-		-
Black (%)						+		ns	+
Immigrants (%)			+			+			
Under 14 (%)				–					
Under 17 (%)	ns				-				
Above 65 (%)	ns	ns				-			
Households without car (%)	-		+	++	+		+		+
Average number of cars in household with children									- -
Income (€)		-	-	–	-	-		-	-
Households with annual income below poverty (%)	+								++
College degree or higher	-								

2.1.2 Built environment

The built environment in this research includes what is in the direct environment of the bus stop, in other words: the catchment area, and what is accessible within a certain amount of time when a bus is taken from that stop. In the literature the first part receives a lot more attention than the second part, as can be seen in table 2. The ones who did address accessibility, were able to determine a positive relationship between the number of residents that can be reached when taking a bus and bus ridership (Guo & Huang, 2020) (Chu, 2004). In other words, for a bus stop it is examined how many residents a bus traveller can reach within a certain amount of time with that particular bus stop as origin. When this number of *accessible residents* increases, the *ridership* increases. The *number of jobs that are accessible* from a stop was found to affect *ridership* positively as well (Dill et al., 2013) (Guo & Huang, 2020) (Chu, 2004) (Chu et al., 2006). But other variables are often only examined for their influence on transit ridership in the catchment areas of stops, such as *hospitals* (strong relationship), *educational buildings*, *restaurants*, *stores* (Guo & Huang, 2020) and *facilities* in general (Dill et al., 2013).

A popular category of variables within built environment is the land-use of the catchment area. *Commercial land use* is found to have a positive effect on ridership (Dill et al., 2013) (Guo & Huang, 2020) (Pulugurtha & Agurla, 2012). The variable *land use mix* has varying results, Dill et al. and Zhao et al. did not find a significant relationship, but Guo and Huang did find a strong relationship. In their turn, Guo and Huang did not find a significant relationship with the variable *residential land-use*, but Kerkman et al. found a strong one. The first contradicting results could be caused by the use of different formulas to calculate the *land-use mix*. The second because Guo and Huang examine *residential land-use* and *population density*, two variables that could have strong correlations.

Furthermore, the infrastructure around bus stops is often examined. The *walk-ability* of a bus stop, which is about if the bus stop is easy and pleasant to reach for a pedestrian, is found to positively affect the number of transit users (Chu, 2004) (Ryan & Frank, 2009). Although the explanatory power turned out to be rather small in the latter study, adding *walk-ability* as an explanatory variable did lead only to an increase in the explanatory power of the model by 0.02%.

The explanatory power of Ryan and Frank is remarkably low, only 32%, although many of the variables they included appeared in different studies and these studies have explanatory powers above the 70%. A reason for this could be that Ryan and Frank did not transform the dependent variable (in section 3.3.2 is explained why this is important). Other independent infrastructure variables that are found to influence transit ridership positively are *street connectivity*, the *length of multi-use paths*, *length of bike-paths* (Dill et al., 2013), *speed limit*, the p*resence of an one-way street* (Pulugurtha & Agurla, 2012), *sidewalks*, *street density* and *highway accessibility* (Zhao et al., 2005).

Table 2: Built environment

	Dill et al. (2013)	Guo & Hang (2020)	Kerkman, Martens & Meurs (2015)	Pulugurtha & Agurla (2012)	Ryan & Frank (2009)	Chu (2004)	Chu (2007)	Cervero, Murakami & Miller (2010)	Thompson (2012)	Zhao et al. (2005)
Std	No	Yes	Yes	No	Yes	No	No	Yes	No	Yes
Population density (#/m2)								++	ns	+
Household density (#/m2)										-
Jobs (#)	+	++				+			+	
Employment density (#/m2)									ns	++
Residents accessibility (#)		+				+				
Jobs accessibility (#)	+	+				+	+			
LU residential (%)		ns	++							
LU commercial (%)	+	+		+						
LU agricultural (%)			-							
LU sociocultural			+							
LU institutional				+						
LU industrial				-						
LU mix (%)	ns	++								ns
Dummy downtown	+									
Dummy university	+	+								
Distance to city centre (km)	-	-	- -							
Distance to station (km)		ns								
Distance to nearest stop (m)								ns	ns	-
Walkability					+	+				
Transit accessibility						+				+
Highway accessibility										+
Street density										+
Speed limit				+						
Parking fees									+	
Hospitals		++								
Restaurants, stores and hotels		+								
Scenic spots		ns								

2.1.3 Level of service

Level of service variables are often examined too, Ryan and Frank (2009) were the only ones to capture *Level of Service* as a whole (LoS). The definition they used was the number of bus routes divided by the average headway time (the time between two consecutive busses). Doing so they found a strong relationship between ridership and LoS. Dill et al. (2013) and Kerkman et al. (2015) treated *headway* as a separate variable and found the more busses arrive per hour, the higher the *ridership*. Moreover the possibility to *transfer to different directions* or to *transfer to a different transit mode* turned out to be important explanatory variables for ridership (Dill et al., 2013) (Kerkman et al., 2015) (Guo & Huang, 2020) as well as the bus stop being located at a *Park and Ride*. (Dill et al., 2013)(Chu, 2004). The existence of *other bus stops in the catchment area* of a specific stop is found to have a negative effect on the bus ridership in two of the four studies (Dill et al., 2013)(Chu, 2004) and non-significant in the other two (Guo & Huang, 2020) (Kerkman et al., 2015). Besides the quantitative side of LoS there is a qualitative side, such as *feeling of safety*, *information availability*, *customer service* and *cleanliness*. These variables were all found to affect ridership by Syed and Khan (2000).

Table 3: Level of service

	Dill et al. (2013)	Guo & Hang (2020)	Kerkman, Martens & Meurs (2015)	Ryan & Frank (2009)	Chu (2007)	Hunt et al (1986)	Cervero, Murakami & Miller (2010)	Thompson (2005)
Std	No	Yes	Yes	Yes	No	No	Yes	No
LOS				++				
Modal transfer stop	+	ns						
Transfer stop	+	++	+					
Transit centre	+		+					
Terminal		ns	+					
Average headway (min)	-		- -				- -	
Directions (#)			+					
Frequency per direction (#3/hour)			++					
Total bus stops within buffer (#)	-	ns	ns	-				
Total bus lines within buffer (#)		-						
Total light rail stations within buffer	-							
P+R	+				+			
P+R parking spaces (#)					+			
Near streetcar stop					+			
On radial routes					+			
Daily feeder trains (#)							++	
Area covered by bus (m2)						+		
Wait time for transit								ns
Dynamic information			+					
Benches			++					

2.1.4 Prices and other variables

In literature there are two variables found that fall under the price category: *gasoline* and *fare prices*. In Washington State in the U.S. Stover and Bae (2011) found the higher the *gasoline prices*, the higher the *bus ridership*. The *fare price* is found to have a negative relationship with *bus ridership* in multiple studies (Sale, 1976) (Liu, 1993). Furthermore, there are variables that could not be categorised such as the *weather conditions* or availability of *shareable e-scooters*. *Humidity*, *wind* and *rain* were found to affect *bus and metro ridership* especially in urban areas. (Zhou et al., 2017). Zuniga-Garcia and Machemehl (2020) found that busses and *shareable e-scooters* are competitive and similar results were found by Luo, Zhang, Gkritza, and Cai (2021). This means that the presence of *shareable e-scooters* has a negative effect on bus ridership.

2.1.5 Conclusion variables

There are four categories of explanatory variables for bus ridership: demographics, built environment, level of service and prices (see figure 1). Every variable that appears in more than one study is at least found once to have a significant effect on ridership. So none of the variables can be disregarded for this research. Furthermore, the variables *residents accessibility* and *jobs accessibility* turned out to be affecting bus ridership significantly, but more variables in terms of accessibility have not been studied before. *Hospitals* have been found to be an important factor in the catchment area of a stop. It would be interesting to examine if the *accessibility of hospitals*, and *other facilities* would have a significant effect on ridership as well.

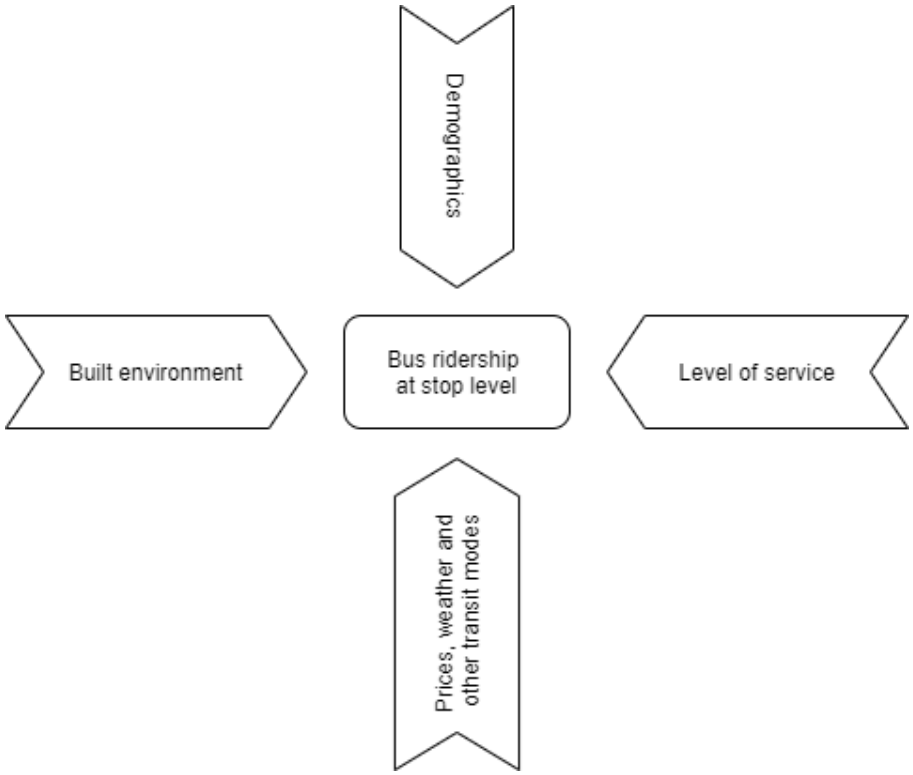

Figure 1: Conceptualisation

2.2 Ridership explaining and prediction methods

In literature several methods have been applied to explain or predict ridership. The advantages and the disadvantages of the methods are displayed in table 4. One of the traditional methods is the four-step model (Cardozo, García-Palomares, & Gutiérrez, 2012). The model is an iterative mechanism which starts with an estimate of the total number of trips that are made in a region based on results of travel surveys conducted among households (McNally, 2007). In the second step the flows between points, called trips, in the region are estimated. Subsequently, it is estimated which transit mode will be used for each trip. The fourth step is to estimate which route travellers will choose for each trip. Busy roads could lead to a difference in a trip is made or not, mode choice and route choice. So the results after the fourth step will be used as feedback for step two, three and step four. The four step model is mostly used for regions or bus routes and it ignores heterogeneous properties within a region or around a route (Chu, 2004).

A method that has grown in popularity the last decade is the direct ridership model at stop level (DRM). DRMs are especially useful for quick, first-cut, order-of-magnitude analyses, according to Upchurch and Kuby (2014). 'Direct' in DRM refers to the transit ridership of a stop being directly calculated by examining local variables (Mucci & Erhardt, 2018). In other words, the influence of variables in the catchment area of the stop on the ridership is estimated (Durning & Townsend, 2015). This makes DRM more suitable for identifying local influences than the more traditional four-step model. According to Chu (2004) another benefit of DRM over the four-step model is the ability to represent competing effects and to represent service variables. Other advantages of DRMs are the high transparency, high explanatory power and the ability to experiment with changes in the transit network (Kuby, Barranda, & Upchurch, 2004).

Fuzzy inference can, just as DRM, be used for predictions at stop level. This method uses multiple if-statements in a row, all including an explanatory variable and a value judgement in terms of high, average or low, to estimate ridership (Kikuchi & Miljkovic, 2001). But according to (Kikuchi & Miljkovic, 2001) this method, does not lead to remarkably better predictions than

a DRM. Moreover, it is easier for DRMs to cope with changes in variables influencing ridership (Chu, 2004).

A structural equation model is a more confirmatory model and is less suitable to use for exploring new variables because alternative models, such as multiple regression, could fit the data equally well or even better (Nunkoo & Ramkissoon, 2012). As one of the objectives of this research is to be able to make a first prediction of ridership at stop level, DRM is the most suitable method. There are different kinds of DRMS and each DRM can deal with different points of attention, which arise when estimating bus ridership.

Table 4: Main ridership models

Method	Used by	Strengths	Weaknesses
Four step model	(McNally, 2007)	Dealing with complexity (McNally, 2007)	Expensive, lack flexibility to deal with changes to a transit network (Gutiérrez, Cardozo, & García-Palomares, 2011) (Chu, 2004)
Fuzzy inference	(Kikuchi & Miljkovic, 2001)	Is able to make use of logical explanatory power and understanding of patterns (Kikuchi & Miljkovic, 2001)	Low sensitivity of predictions to changes in variables affecting ridership (Chu, 2004)
DRM	See table below	High transparency, high explanation power and the ability to experiment with changes in the transit network (Kuby et al., 2004). Useful for evaluating exploratory variables (Nunkoo & Gursoy, 2012)	Causal relationships are unknown (Kikuchi & Miljkovic, 2001)
Structural equation model	(Sohn & Shim, 2010)	Able to identify indirect or cyclic relationships among variables (Sohn & Shim, 2010). Takes measurement errors into account (MacKenzie, 2001)	Complexity, ambiguity and is less suitable to explore new variables (Nunkoo & Ramkissoon, 2012)

2.3 DRM models

Within DRM there are several different models that could be applied, some of them are displayed in table 5. The models mentioned in the table, are all regression-base. There are three main purposes of using regression analysis: explaining, predicting and describing. According to Shmueli the first purpose is about retrieving insights in causal relationships between the dependent variable and the independent variables. The second one is used to predict new or future observations. Describing is about summarizing data and in comparison with explaining, it focuses less on causal inference and the focus is more on displaying data structures in a compact manner (Shmueli, 2010). The different purposes all have different requirements. We will have a closer look at the differences between explaining and predicting, while examining important features which should not be overlooked when bus ridership is explained or predicted.

2.3.1 Endogeneity and omitted bias

Because the focus with explaining is on causality, the simultaneity bias is important to take into account. Simultaneity bias is a form of endogeneity and it means that the outcome variable of the regression analysis is a predictor of an explanatory variable instead of a response to the explanatory variable (Lynch & Brown, 2011). An example: when the presence of a digital travel information display is used to predict the number of travellers by bus at stop level, it is likely that the digital display is placed because of the high number of travellers instead of the display attracting many travellers (Kerkman et al., 2015). In this case, the digital display is an endogenous variable. The endogeneity bias plays a big role for most of the variables mentioned above. Another example is

transit companies adjusting their bus schedules based on the expected ridership. For instance if they expect many travellers will make use of a stop. A way to deal with endogeneity is the Two Stage Least Squares test, but the data requirements are high (Shepherd, 2009). A less refined way is to exclude to variables that are expected to cause an endogeneity problem (Chu, 2004).

Fortunately, when predicting is the goal the focus should purely be on getting as high a predictive power as possible and it is unnecessary to examine the causal relationships between the independent variable and the dependent variables (Shmueli, 2010). Allison (2014) argues, that because the focus is purely on the predictive ability, the omitted bias can also be disregarded. Omitted bias occurs when the influence on the dependent variable of an unobserved variable is attributed to observed variables. This will result in biased values for the importance (coefficients) of the independent variables (an example is given in the appendix, section 9.1). According to Allison, for predicting it does not matter if the coefficients are not the true value, as long as all the coefficients together can be used to calculate optimal estimates.

2.3.2 Non-stationary relationships

Important terms for both explaining and predicting are spatial auto-correlation and non-stationary relationships. The first one is about the common knowledge in the field of Geography that values close to each other are more equivalent than those further away (Tobler, 1970). The latter means that relationships between the independent variable and the dependent variables can vary based on their location (Murack, 2013). The regression models that can deal with spatial auto-correlation and non-stationary relationships are called spatial regression models. According to Murack, no spatial regression model can take both points into account. Non-stationary can be overcome by using local regression models, such as Geographically Weighted Regression (GWR) models (Propastin, Kappas, & Erasmi, 2008). The local ridership models show better results than the global models (Chiou, Jou, & Yang, 2015) (Gan, Feng, Yang, Timmermans, & Luo, 2019) (Chow et al., 2006), but a downside is that the results of a local study cannot be extrapolated to areas that are not the main study area (Propastin et al., 2008). The aim of this study is to develop a prediction model for the Netherlands, with as study area concession area Arnhem-Nijmegen. Therefore, global regression models will be further looked into.

2.3.3 Spatial auto-correlation

The spatial lag model (SLM), spatial lag X model (SLX) and spatial error model (SEM) are three global models which can deal with spatial auto-correlation, also known as spatial dependence (Anselin & Bera, 1998). The first model examines if the value of the dependent variable is affected, besides by its own independent variables, by the value of the dependent variable of its neighbour (neighbour is a broad concept, see section 3.5.1). For the SLM model the neighbouring independent and dependent values should be known. The objective of this study is to find out if it is possible to make predictions for regions in the Netherlands, for which no ridership data is available, based on the regression results of Arnhem-Nijmegen. Thus, for these regions the values for the neighbouring independent variable are unknown, which makes the SLM model impractical for this study.

For SLX the dependent variable is influenced by its own independent variables and its neighbours independent variables. The independent variables are selected based on their potential influence and on their nation-wide availability. So for this model, neighbours can be regional. The spatial error-model deals with auto-correlation in the error term, thus it examines if there are regions with remarkably high or low residuals and bases its estimates on this. To retrieve the error term, first the observed value and the predicted value should be known, as this will not be the case for other parts in the Netherlands. The SLX model is the only method suitable for prediction. But the SEM model can still be important in this research. If spatial auto-correlation exists in the error terms and a normal OLS model is estimated, it could lead to wrong significance values for variables and it could lead to a sub-optimal prediction (Anselin & Bera, 1998).

For predictions in the Netherlands, SLX is the most suitable model. For explaining the bus ridership in Arnhem-Nijmegen both the SLX and SEM could be used, as it will depend heavily on the data which model will perform better. For instance, Gan et al. (2019) applied the spatial lag model and the spatial error model and found that the latter was the best performing one. This does not mean the spatial error model will be the best model for all studies trying to explain ridership, as it only means that the model of Gan et al. wasn't able to include the explanatory variables in the model that cause the spatial auto-correlation (Kerkman et al., 2018).

2.3.4 Count data

Pulugurtha and Agurla (2012) argues that ridership numbers are count values. An OLS model assumes that the dependent variable is continuous (Peel, Goode, & Moutinho, 1998), meaning that the variable could be any number. But count values are bound at zero, for example a bus stop cannot have a negative number of travellers. Violating the continuous dependent variable assumption could lead to predictions with a negative value (Peel et al., 1998). Pulugurtha and Agurla (2012) point out that the Generalized Estimating Equation (GEE) can, in contrast to OLS, deal with count data and therefore they consider GEE as a more suitable model for explaining and predicting bus ridership.

A downside of GEE is that there is not a universal goodness-of-fit test (Ballinger, 2004) and although ridership can indeed be seen as count values, treating it as such is found to not have a particular advantage for predicting bus ridership when ridership is transformed (Dill et al., 2013). In section 3.3.2, more is described on transformations of variables.

Table 5: DRM models

Method	Used by	Strengths	Weaknesses
OLS	(Kerkman et al., 2015) and more	Low cost and quick response (Gan et al., 2019) Can be used for explaining (Kerkman et al., 2015)	Susceptible for multicollinearity endogeneity bias, spatial dependence and spatial heterogeneity (Guo & Huang, 2020) (Gan et al., 2019)
Spatial lag model / spatial auto regression	(Gan et al., 2019)	Able to deal with auto-correlation between the independent variable and the associated spatially lagged variable (Anselin & Bera, 1998)	Can't deal with spatial error auto-correlations (Anselin, 2003). If spatial point has no neighbours, it becomes a non-spatial formula (Goulard, Laurent, & Thomas-Agnan, 2017)
Spatial lag X model (SLX)	No one	Can deal with spatial auto-correlation between dependent and neighbouring independent variables	Can't deal with spatial error auto-correlation
Spatial error model	(Gan et al., 2019)	Able to deal with auto-correlation between error terms (Anselin & Bera, 1998)	Can't deal with spatial lag auto-correlations (Anselin, 2003)
Generalized estimating equations framework	(Pulugurtha & Agurla, 2012)	Is able to deal with count data (Peel et al., 1998)	There is no universally accepted goodness-of-fits test and the predictability cannot be easily assessed (Ballinger, 2004) (Davarzani, Peeters, Smirnov, Karel, & Brunner-La Rocca, 2016)
2-stage least squares test	(Taylor, Miller, Iseki, & Fink, 2003)	Able to deal with endogeneity (Taylor et al., 2003) (Shepherd, 2009)	Excessive data requirements (Shepherd, 2009)

2.3.5 Ridership prediction in literature

Many of the studies described above focus only on the explaining part of ridership, Cervero et al. (2010) were one of the few that looked into the prediction capability. The study contained 69 Bus Rapid Transit (BRT) stops and they concluded DRM was suitable for predicting ridership at low ridership stops and at high ridership stops. Their observations/independent variable ratio is relatively low, with 9 variables for 69 observations. This could be an explanation of their remarkably explanatory power of 95.2%.

Guo and Huang (2020) predicted the monthly ridership for 72 Mass Rapid Transit (MRT) stops. Their predictions can be considered as accurate for 80% of the stops. The absolute relative error (first the observed value is subtracted from the predicted value and the obtained value is divided by the observed value) is lower than 50%. But their prediction intervals, intervals for which the model can predict with a 95% certainty the observed ridership will belong to are quite wide.

For instance when a monthly ridership of 10,000 is observed, the lower bound of the prediction interval is around 3,000 and the upper bound is approximately 17,000.

2.3.6 Conclusion models

Existing literature on explaining and predicting bus ridership at stop level shows promising results, with the first being more often examined than the latter. For instance, the explanatory power is tested in the Netherlands, but the prediction ability is not (Kerkman et al., 2015). And when the prediction ability is evaluated, the sample sizes are small. This research will try to predict bus ridership numbers at stop level for the concession area Arnhem-Nijmegen, by examining demographics, built environment and level-of-service variables, which ones exactly will be discussed in the next chapter. For the predictions, a Direct Ridership Model is the most suitable model, since it can be quickly applied and it can capture local influences. It turned out to be important that the DRM is global, since local DRMs cannot be extrapolated outside the study area. This means the model won't be able to deal with non-stationary relationships.

The OLS model is commonly used and has proven itself to be sufficient to explain bus ridership. But to get more accurate prediction spatial auto-correlation can be taken into account. For this research it is likely that the independent variables affect the neighbouring dependent variable, when a stop has more directions than its neighbours, it is likely that this has a negative effect on the bus ridership of the neighbouring stops. The SLX model, not used before for DRMs, can capture this effect and thus this may result in more accurate predictions than the OLS. Furthermore, it is improbable that all of the ridership will be explained by variables that will be included in the model. What cannot be explained will be captured in the error terms and in those terms there is the chance of spatial auto-correlation. To deal with spatial auto-correlation, the spatial lag x model and the spatial error model will be applied in this research for explaining. The OLS and SLX will be used for predicting.

3 Method

In chapter 2 we found that demographics, built environment, level of service and prices are known categories to influence bus ridership. This chapter will discuss which variables are included in the model and which are not and will provide the reason behind it. Afterwards, the operationalisation of the variables is explained and the data is examined. Then the OLS and the spatial models will be clarified, after which an explanation of prediction accuracy follows. But first, the study area Arnhem-Nijmegen is described.

3.1 Study area

The model will be based on the concession area Arnhem-Nijmegen, see figure 2 (picture is edited, original is from (Regenmaker, 2019). The area belongs to one of the five biggest urban agglomerations in the Netherlands. The largest cities in concession Arnhem-Nijmegen are unsurprisingly, Arnhem and Nijmegen. Nijmegen is with almost 180,000 residents the tenth largest city in the Netherlands and is closely followed by Arnhem. Arnhem with 160,000 residents is at the thirteenth place (Allecijfers.nl, n.d.-b). The cities are the economic heart of the region and are located 15 km apart from each other (Kerkman et al., 2015). The concession area exists of 17 municipalities and has a variety of different kinds of urbanization, there are small cities, such as Doetinchem (58,000) and Duiven (25,000) and plenty of small villages. The population of the concession area is just below 740,000 residents in 2019 and the area covers an area of more than 1100 square kilometers (Eurostat, n.d.), which makes the population density 670 per km2.

Arnhem-Nijmegen has one university, the Radboud University in Nijmegen and the area has two major tourist attractions: a zoo and an open air museum. The area is connected by five highways and the area is easily accessible by train, as it is possible to go into every direction by train (Gemeente Arnhem, 2016). There are 23 train stations and the train is, together with the bus, the only public transport mode in the region. There are 988 bus stops in the concession area, figure 3 shows how the stops are divided over the concession area and it shows the number of rides per bus stop. The total number of city lines and region lines is 52 (Open mobility data, 2019) and six of those lines are run by trolley busses in Arnhem, the only trolley busses in the Netherlands (Arnhem2day, n.d.).

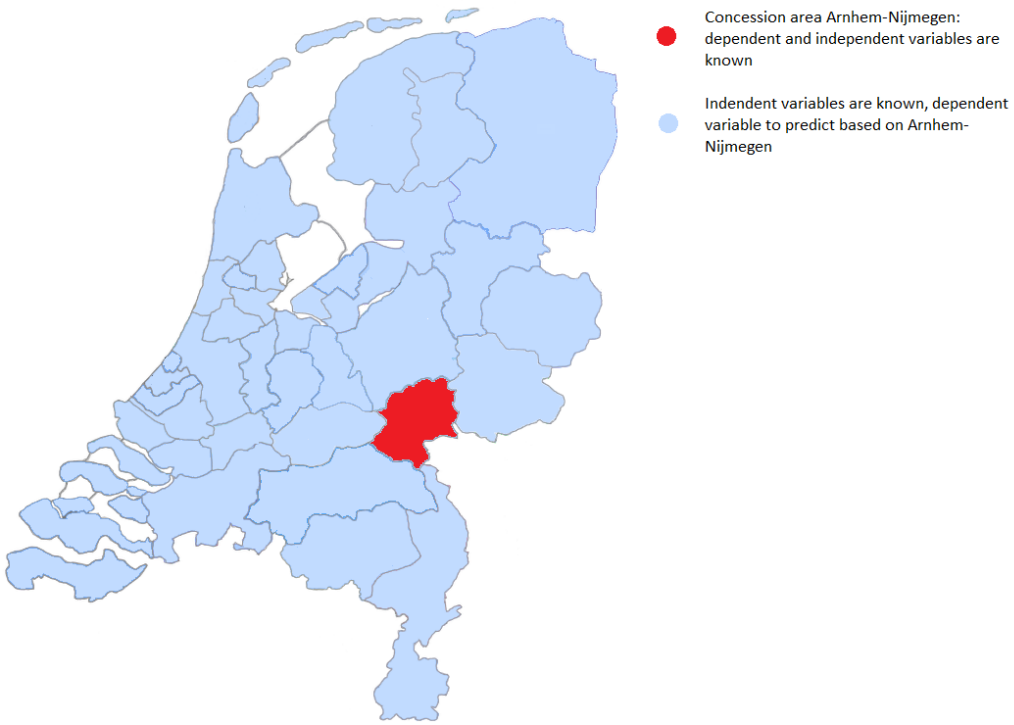

Figure 2: Map of the Netherlands, study area is highlighted

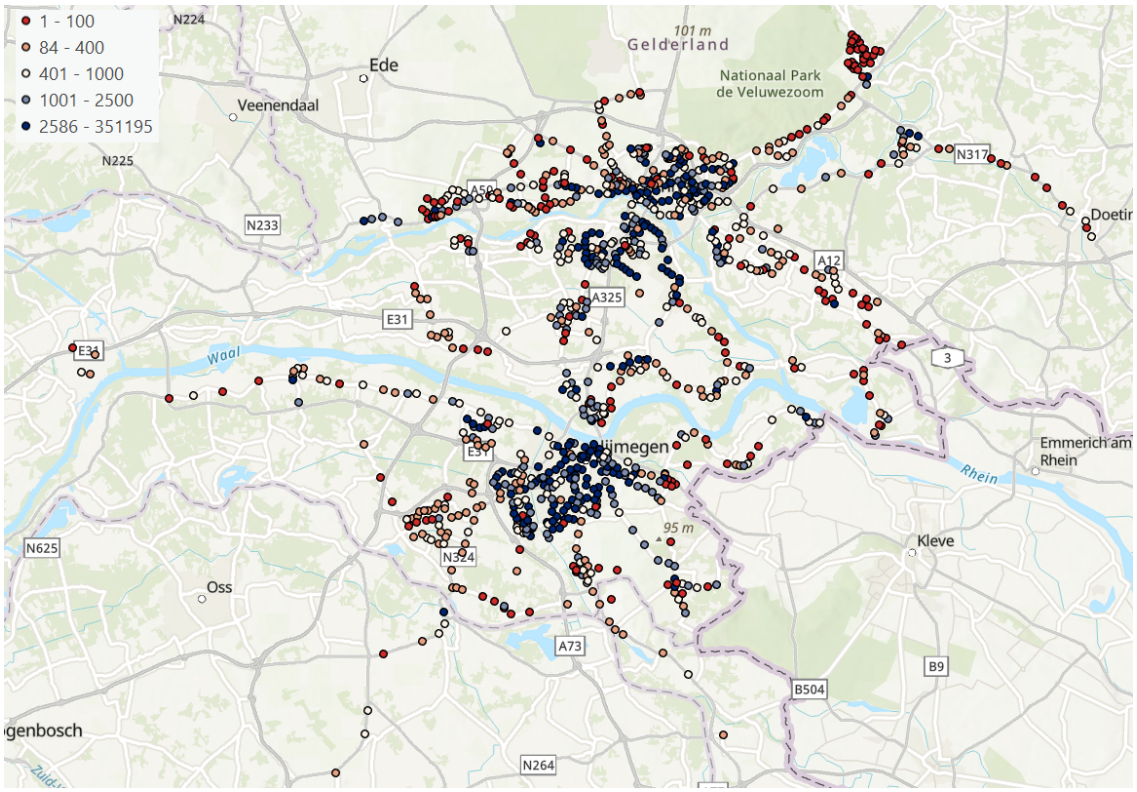

Figure 3: Bus stops and number of rides in concession area Arnhem-Nijmegen

3.2 Variables and data

In Chapter 2 variables were identified that were examined in scientific literature. That literature study showed that many variables were interesting to be included in this research. The starting point was to include as many as possible, as long as the data required for the variables was free of charge and available nationwide. Most of the data for this study is provided by public instances, such as Centraal Bureau voor de statistiek (CBS), Dienst Wegverkeer (RDW) and Dienst Uitvoering Onderwijs (DUO). Furthermore, Movares and open data websites such as the NDOV Loket, OpenMobilityData and Open State Foundation are used to retrieve data as well.

3.2.1 Rides

The most important variable of this study is the dependent variable *rides*, the variable that needs to be explained by looking at other variables. The data for *rides* is retrieved from smart card data provided by OV Oost and it contains the amount of rides have been made between two specific stops per hour on all the week days in November 2019. For instance, the data shows that in November 2019 a total of 141 travellers departed between 18:00-19:00 from Arnhem central station and got off the bus at Arnhem Velperplein. According to Chu (2004) the best predictor for the number of arriving travellers at stop level is the number of departing travellers at the same stop and vice versa. After inspecting the smart card data, it turned out that there were more known origin *rides* than known destination *rides*. The difference between known origins and known destinations is caused by travellers forgetting to check-in or to check-out. Because the number of known origins is higher, this study will work with departing travellers as bus ridership. Bus stops can consist of multiple quays, all the departures of the different quays but at the same stop will be summed and this will be seen as bus ridership at stop level and will be labelled as *rides*. The number of *rides* per stop on working days will be expressed per month.

3.2.2 Demographics

For many of the in chapter 2 identified demographics that influence bus ridership, such as *gender*, *age*, *ethnicity*, *household size*, *income* and *car ownership*, data could be found on the CBS website. CBS works with three different scale levels based on zip codes: zip code 4, zip code 5 and zip code 6. The number refers to the number of characters in the zip code. A Dutch zip code contains four numbers and two letters. A zip code 6 exists of the full zip code and thus gives the most detailed information. Unfortunately, as CBS protects the privacy of residents, when there are too few residents for a zip code the data for that zip code will not be made public. This resulted in many empty values for the independent variables. Therefore, the choice has been made for the zip code 5 scale-level.

To calculate the values for the demographics of a bus stop, a catchment area, with a radius of 400 meters is taken as this is the maximum distance most travellers are willing to walk to a bus stop (van der Blij, Veger, & Slebos, 2010) and it fits in the ambition of Arnhem and Nijmegen for all their residents to have at least one bus stop within 400 meters (Omroep Gelderland, 2019). A catchment area of a stop can consist of multiple zip codes. With the use of ArcGIS Pro, per stop it is examined which zip codes are located in the catchment area and for what percentage does this area consist of a specific zip code. This way, the average values for demographics in the catchment areas could be calculated. An example of the calculation for the percentage of *youth* for a catchment area can be found in the section 9.3. It can happen that multiple catchment areas overlap, how this is dealt with is described in the built environment section.

Two age variables are included, one is the percentage of *youth*, 15-24 year olds and the other is the percentage of *elderly* (65+). This way students with their student travel product, which allows for cost-free travelling and people who are retired or close to retirement are included separately in the model, leaving children and mostly working people not to be named specifically.

In chapter 2 two kinds of variables for car ownership were discussed: *average car ownership* and *households without a car*. The first appeared in one study and was found not to be significant, the latter appeared statistically significant in five studies. For the Netherlands, data on *average*

car ownership and on *average car ownership per low, middle and high income* was available. *Car ownership for low incomes* says probably more about households without a car, than *car ownership for high incomes*. But again, the more detailed the information is, the more empty values are found. So, only *average car ownership* is included in the model. For income, the lower incomes can be included, but only on a zip code 4 scale. In the model the percentage of *households with low income* is included, which is the percentage of all households that have an income that belongs to the 40% lowest incomes.

Politics are not included in the studies discussed in chapter 2, but the political preferences can say something about how they think and about their travel behaviour. Van der Groot (2012) examined what the Second Chamber would look like based on the respondents of her research, who are frequently public transit users. For the parties PvdA, SP, Groen-Links, D66 and Partij voor de Dieren it would mean an increase in seats. For the VVD, PVV and CDA it would lead to a decrease of seats. Van der Groot does not comment on all the parties which are currently present in the Second Chamber since new parties have joined the Second Chamber after the elections of 2021. Above mentioned parties are good for 122 of the 150 seats in the current Second Chamber (Tweede Kamer, 2021). Because of this distinction the choice is made to include politics in the model. The PvdA, SP, Groen-Links, D66 and Partij voor de Dieren will be clustered together under the label 'green parties'. The VVD, PVV and CDA will be clustered too and the other parties will be disregarded for this research. Data for the Second Chamber elections of 2021 is made available by data.overheid.nl. With the use of the Python script created by Kloosterboer (2021) the xml files of the website are converted to csv files. Whereafter the percentage of *green votes* was linked to data on the location of the polling stations retrieved from the Open State Foundation (n.d.). To receive information on how people vote around the bus stops, every stop is linked to the closest polling station.

3.2.3 Built environment

In chapter 2 three groups within the built environment variables were described: within the catchment area of the bus stop, what can be reached while taking the bus from a bus stop and the walk-ability. First, the within catchment area will be discussed.

Six land-use variables will be studied, with data obtained from CBS (RHDHV, 2019). The six variables are *residential*, *social and commercial* (consisting of retail, hotels, bars and restaurants), *business* (offices, industrial and public facilities), *agricultural*, *recreational* (parks and sport facilities) and *amusement* (amusement parks and zoos). The first three variables were found to be affecting ridership the most in chapter 2. The fifth variable has not been studied yet in this composition, the number of parks turned out to individually influence bus ridership negatively (Dill et al., 2013) and sport facilities were not examined at all. But because it is assumed people will not take the bus often to sport facilities, the two are merged together into one variable. *Institutional land use* appeared in one study and did have a significant relationship (Pulugurtha & Agurla, 2012). But it will not be included in this model, since the assumed most influential institutional facilities will be examined individually.

Institutional facilities that will be included are *hospitals*, *higher education* (HBO and WO) and *secondary education*. The latter is not examined before, but it is included since it appeared from 1.1.1 that half of the kilometers travelled by bus, tram and metro was made by pupils and students. *Hospitals* are examined, since of all the facilities studied, such as *hotels* or *scenic spots*, by Guo and Huang (2020), *hospitals* turned out to be the most influential. Other facilities that will be studied are *park and rides*. The data for the locations of *hospitals* is provided by Movares and the locations for *educational institutions* are retrieved from DUO (2021). The RDW (2021a) provided the information for the *park and rides*.

A problem that occurs in the denser areas, is that bus stops have overlapping catchment areas. For the catchment areas with overlap, the values for demographics are calculated the same way as for non-overlapping catchment areas. But to compensate variables being counted twice, the variable number of *other bus stops within the catchment area* is included in the model. It is expected this variable will have a negative influence on bus ridership at stop level. It should be

noted that this method could still lead to overlapping catchment areas even when the data says there are not any other bus stops within the catchment area, because catchment areas of two bus stops can overlap, without being in the catchment area of the other stop. To prevent bus stops from being seen as competitive if they only share a small area only the actual number of *other bus stops within a catchment area* are counted. A visualization can be found in appendix 9.5.

Overlap is a problem for facilities as well, there are cases where a *secondary school* is in three catchment areas. This could lead to underestimating the importance of *secondary schools* on bus ridership, as it is likely that pupils will go to the nearest stop. This results in other stops having a *secondary education* facility in their catchment area, but will not see a higher number of rides because of it, which causes the coefficient for *secondary education* to decrease in strength. To prevent this a new built environment category will be added, namely how many times a bus stop is the closest bus stop to a facility. For instance, for every *secondary school* the nearest stop is identified. The value for the variable *closest to secondary school* will be one for that bus stop. If that bus stop is the closest bus stop for two *secondary schools*, the value will be two.

There are multiple reasons for including facilities *within the catchment area* and how many times a stop is the *closest stop to a facility*. It could be that facilities have multiple entrances with each a bus stop. Or it could be that, for instance, pupils do go to a bus stop which is not the nearest, since they need a particular bus line. Furthermore, for *higher education* it could be useful as well, because only data on one single location is known per university and a campus often exists of multiple bus stops. So, by including *higher education within catchment* areas it is examined if the effect of this data limitation can be reduced. In the next chapter, the independent variables will be estimated per category to see their influence on bus ridership. During this step the Forward Selection process is applied, see section 3.4.1, to identify per facility whether the *within buffer* or *closest to* variant is affecting bus ridership the most. For clarity reasons, the variant which has the least influence on bus ridership will be disregarded from the model.

For *hospitals*, *higher education*, *secondary education* and *park and rides* the accessibility will be studied as well. Accessibility has been studied for residents and jobs and it showed significant results, but including accessibility for these facilities is new. The accessibility could also have been examined for other variables, such as how many *elderly* can somebody access from a certain bus stop, but *hospitals*, the *educational facilities* and *park and rides* are assumed to be the most promising variables.

For the accessibility variables it is determined with the use of the Verbindingswijzer (VBW), how many *jobs*, *residents* etc. can be reach from a bus stop within 45 minutes, without transfers. The time-limit is set to 45, because Gimenez-Nadal and Molina (2014) found this is the average commuting time in the Netherlands and for consistency it is applied to the other variables besides jobs as well. The number of transfers was set to zero, to purely look at what a traveller can reach from a stop. If transfers were allowed, the area a traveller could reach would grow significantly, which makes it harder to find significant relationships. To calculate the accessibility variables, data required for the within catchment area variables are loaded in the Verbindingswijzer which is a tool that can retrieve the accessibility variables. More on the Verbindingswijzer is in appendix 9.2. The *accessibility of train stations* has been left out of the model, as it turned out there are only 17 stops out of the 965 in the area where a traveller cannot reach a station within 45 minutes.

Another built environment variable is *parking costs* around a bus stop. Different data sets retrieved from the RDW (2021b) were merged, to be able to calculate the weighted average parking cost per hour.

Unfortunately, no data could be found on city centres and because data should be available nation wide, retrieving the data by hand is considered to be too time consuming. Thus, the variable *distance to nearest city centre* will not be included in this model. *Distance to station* appeared once in a study and was found not to be significant, but this study did contain the distance to city centre as well (Guo & Huang, 2020). This study will look into the effect of *distance to station* on ridership, when it is the only distance variable in the model. Variables for *walkability* also could not be found but it appeared from Ryan and Frank (2009) the explanatory power for ridership was limited.

3.2.4 Level of service

In chapter 2 it appeared that frequency and whether a transfer can be made on a stop are important for explaining ridership and are therefore included in the model. The latter is expressed in the number of directions a traveller can go from a stop. Furthermore, several physical bus stop characteristics are included, namely the availability of a shelter, bench, digital information display, bike stands and if the bus stop is accessible to the disabled. The problem with endogeneity could be severe with these variables and therefore they are less suitable for prediction. But they are included, since if there is a strong relationship between a characteristic and ridership, it could be an indicator for a missing explanatory variable.

3.2.5 Data limitations

Besides the lack of specific data there are other data limitations, for instance the time window of the data. The ridership data is about November 2019, but the oldest data set comes from 2016. It is unfortunate the years do not match, but demographics of a neighbourhood do not rapidly change (Zwiers, 2018). There is a risk of newly developed neighbourhood finishing between 2016 and 2019. This will cause high ridership for close bus stops, which cannot be explained. If this is the case, these outliers can be easily identified and removed when the first results are examined.

Variables such as weather, ticket fare and qualitative level of service will not be included. The weather can be used to explain bus ridership, but is less suitable to predict bus ridership and there is not any variation in the ticket fare nor qualitative level of service in the concession area, so these cannot be used to predict bus ridership. Shareable scooters were not available in the concession area the year this research examines (2019) so this variable also is not included.

3.2.6 Descriptive statistics per bus stop

Below, an overview is presented of all the variables that are included in the first model. The minimum, mean and maximum values of the 965 stops are displayed for every variable. There exists large variance in the dependent variable and in the independent variables related to residents, jobs, frequency and the address density. The standard deviations for these variables are high and for some of them the standard deviation is greater than the mean.

Most within buffer- and closest to variables seem to behave like a dummy variable, a variable which only can have the value zero or one.

Of the facilities, there are only multiple secondary schools within a catchment area.

Table 6: Descriptive statistics

	Unit	Mean	Minimum	Maximum	Std. Dev
Rides	#/month (week days only)	2732.508	1	351195	14198.615
Residents	#	1439.724	0.779	6640.136	1142.312
Average household size	#/household	2.171	1.238	4.965	0.369
Average car ownership	#/household	0.957	0.21	1.53	0.256
Female	%	50.574	26.729	56.288	2.239
Youth (15-24)	%	12.712	4.242	50.336	5.557
Elderly (65+)	%	20.089	1.737	53.912	7.914
Households with low income	%	40.622	14.01	70.208	11.795
Residents Western background	%	9.981	0	26.545	2.299
Residents Non Western background	%	8.787	0	57.476	10.115
Green votes	%	54.055	28.329	83.013	12.514
LU residential	m2	45.53	0	96.619	27.768
LU business	m2	1.945	0	92.557	8.768
LU social and commercial	m2	2.476	0	59.752	6.177
LU agricultural	m2	1.762	0	96.791	10.570
LU recreational	m2	3.69	0	64.138	8.447
LU amusement	m2	0.407	0	44.867	3.165
Address density	#	4688.102	0.41	44356.647	6405.905
Distance to station	km	6.514	0.006	38.395	6.221
Parking rate	€/hour	0	0	0.042	0.003
Jobs within buffer	#	349.497	0	13796	1057.926
Train stations within buffer	#	0.054	0	1	0.226
Hospitals within buffer	#	0.015	0	1	0.120
Higher education within buffer	#	0.011	0	1	0.106
Secondary education within buffer	#	0.163	0	3	0.462
P+R within buffer	#	0.026	0	1	0.159
Other bus stops within buffer	#	1.742	0	8	1.122
Closest to train station	Dummy	0.024	0	1	0.153
Closest to hospital	Dummy	0.005	0	1	0.072
Closest to higher education	Dummy	0.004	0	1	0.064
Closest to secondary education	Dummy	0.061	0	3	0.280
Closest to P+R	Dummy	0.009	0	1	0.096
Residents accessible	#	84269.945	21	354560	56437.611
Jobs accessible	#	46531.58	0	122441	23871.410
Hospitals accessible	#	0.798	0	3	0.806
Higher education accessible	#	1.101	0	6	0.833
Secondary education accessible	#	8.422	0	35	7.032
P+R accessible	#	1.537	0	6	1.374
Frequency	#/hour	4.915	0.16	79.47	5.751
Directions	#	2.769	1	36	2.937
Shelter	Dummy	0.634	0	1	0.482
Bench	Dummy	0.662	0	1	0.473
Digital information display	Dummy	0.258	0	1	0.438
Bicycle stand	Dummy	0.275	0	1	0.447
Accessible for disabled	Dummy	0.703	0	1	0.457

3.2.7 Operationalisation

Most of the data cleaning and data preparing for ArcGIS Pro and VBW is done with the use of Python, the scripts can be found in 9. The flow of information is displayed in the operational conceptualisation, in appendix 9.6. The variables, their units, expected sign, source, tool used and the how they are calculated are displayed in table 7.

Table 7: Operationalisation

Variable	Year	Expected sign	Data source	Tool used	Operationalisation
Rides	2019		OV Oost		The number of boarders on a bus stop in November, disregaring the direction
Residents	2017	+	(CBS, 2017)	ArcGIS Pro	Catchment area
Average household size	2017	+	(CBS, 2017)	ArcGIS Pro	Catchment area
Average car ownership	2019	-	(CBS, 2019)	ArcGIS Pro	Catchment area
Female	2017	+	(CBS, 2017)	ArcGIS Pro	Catchment area
Youth (15-24)	2017	+	(CBS, 2017)	ArcGIS Pro	Catchment area
Elderly (65+)	2017	-	(CBS, 2017)	ArcGIS Pro	Catchment area
Households with low income	2018	+	(CBS, 2018)	ArcGIS Pro	Catchment area
Residents Western background	2017	+	(CBS, 2017)	ArcGIS Pro	Catchment area
Residents Non Western background	2017	+	(CBS, 2017)	ArcGIS Pro	Catchment area
Green votes	2021	+	(Kiesraad, n.d.) & (Open State Foundation, n.d.)	Python (Kloosterboer, 2021)	Every stop is linked to the nearest polling station
LU residential	2015	+	(RHDHV, 2019)	ArcGIS Pro	Catchment area
LU business	2015	+	(RHDHV, 2019)	ArcGIS Pro	Catchment area
LU social and commercial	2015	+	(RHDHV, 2019)	ArcGIS Pro	Catchment area
LU agricultural	2015	-	(RHDHV, 2019)	ArcGIS Pro	Catchment area
LU recreational	2015	-	(RHDHV, 2019)	ArcGIS Pro	Catchment area
LU amusement	2015	+	(RHDHV, 2019)	ArcGIS Pro	Catchment area
Address density	2017	+	(CBS, 2017)	ArcGIS Pro	Catchment area
Distance to station	2017	-	(CBS, 2017)	ArcGIS Pro	Catchment area
Parking rate	2021	+	(RDW, 2021a)	ArcGIS Pro	Total parking costs for November, divided by 720h
Jobs within buffer		+	Movares	VBW	Within catchment area
Train stations within buffer	2021	+	(NDOV, n.d.)	ArcGIS Pro	within catchment area
Hospitals within buffer		+	Movares	ArcGIS Pro	within catchment area
Higher education within buffer	2021	+	(DUO, 2021)	ArcGIS Pro	HBO and WO within catchment area
Secondary education within buffer	2021	+	(DUO, 2021)	ArcGIS Pro	within catchment area
P+R within buffer	2021	+	(RDW, 2021a)	ArcGIS Pro	within catchment area
Other bus stops within buffer	2021	-	(NDOV, n.d.)	ArcGIS Pro	within catchment area
Closest to train station	2021	+	(NDOV, n.d.)	ArcGIS Pro	Every facility is linked to the closest stop
Closest to hospital		+	Movares	ArcGIS Pro	Every facility is linked to the closest stop
Closest to higher education	2021	+	(DUO, 2021)	ArcGIS Pro	Every facility is linked to the closest stop
Closest to secondary education	2021	+	(DUO, 2021)	ArcGIS Pro	Every facility is linked to the closest stop
Closest to P+R	2021	+	(RDW, 2021a)	ArcGIS Pro	Every facility is linked to the closest stop
Residents accessible	2019	+	Movares	VBW	Accessible in 45 min without transfers
Jobs accessible		+	Movares	VBW	Accessible in 45 min without transfers
Hospitals accessible		+	Movares	VBW	Accessible in 45 min without transfers
Higher education accessible	2021	+	(DUO, 2021)	VBW	Accessible in 45 min without transfers
Secondary education accessible	2021	+	(DUO, 2021)	VBW	Accessible in 45 min without transfers
P+R accessible	2021	+	(RDW, 2021a)	VBW	Accessible in 45 min without transfers
Frequency	2019	+	(Open mobility data, 2019)	Better Bus Buffer (ArcGIS)	All departing busses on 4-11-2019 divided by 24h
Directions	2019	+	(Open mobility data, 2019)	Rstudio	If available at at least one quay, value is 1
Shelter	2021	+	(NDOV, n.d.)		If available at at least one quay, value is 1
Bench	2021	+	(NDOV, n.d.)		If available at at least one quay, value is 1
Digital information display	2021	+	(NDOV, n.d.)		If available at at least one quay, value is 1
Bicycle stand	2021	+	(NDOV, n.d.)		If available at at least one quay, value is 1
Accessible for disabled	2021	+	(NDOV, n.d.)		If available at at least one quay, value is 1

3.2.8 Missing values

As described the CBS protects the privacy of residents and because of this 9,000 of the 24,000 zip codes have one or more missing values for one of the demographics variables. Luckily, catchment areas of stops exists most of times out of multiple zip codes. Thus, if one zip code with empty values was removed from the data set, it did not lead to empty values for demographics for a bus stop. For demographics which are represented in percentages, such as the percentage of the total population that is between 15 and 24, it is assumed the percentages for the deleted demographics of the deleted zip code are similar to the other zip codes in the catchment area. By removing a zip code, the number of residents cannot be compensated by the other zip codes, but the influence of number of residents will be limited as the CBS protects the privacy if there are less than five residents.

Bus stops with missing values for other variables, such as frequency, were removed from the model. This resulted in removing 23 bus stops out of the 988.

3.3 Model assumptions

3.3.1 Multicollinearity

As the first model is used to retrieve a better understanding what the relationships are and what explains bus ridership, one of the first steps of an OLS-regression is to check for multicollinearity. For predictive purposes, multicollinearity does not have to be examined (Morris & Lieberman, 2018). If there is multicollinearity, it means that some of the independent variables are highly correlated and it could lead to coefficients being assigned the wrong sign. An indicator per variable for multicollinearity is the variance inflation factor (VIF) (C. G. Thompson, Kim, Aloe, & Becker, 2017). The lower the VIF, the lower the multicollinearity. According to Marquaridt (1970) a VIF higher than 10 is an indication of serious multicollinearity. With SPSS the VIF values are retrieved and are displayed in table 9.7 9 and it can be seen that residents, address density, accessibility residents and accessibility secondary education are the only variables with a VIF higher than 10. When the two variables with the highest VIF are removed, residents and accessibility residents, the VIF values for address density and accessibility secondary education drop below 10, see table 9.7 in the appendix.

3.3.2 Outliers and skewness

Outliers in data sets are data points that are far outside the norm for a variable (Stevens, 1984). Osborne and Overbay (2004) names three unfavourable effects of outliers. They can decrease the power of statistical tests and they can enlarge the error variance. Moreover, outliers could bias estimates and outliers could lead to violation of the assumption that residuals (the observed value - predicted value) are normally distributed. One way to identify outliers is to look at the z-score. The Z-score shows how many standard deviations an observation is from the mean (Aggarwal, Gupta, Singh, Sharma, & Sharma, 2019). An observation is seen as an outlier, when its z-score exceeds 3 (Sincich, 1993). Looking at mean and standard deviation statistics in table 6, bus stops with more than 45,000 rides should be labelled as outliers, which would lead to 11 outliers in the data set. Excluding outliers is a common way to deal with outliers, since this will result in the most honest estimates for the relationships between the explanatory variables and bus ridership for the population (Judd & Kenny, 1981).

But by removing the top eleven highest ridership numbers, the model will be incapable to estimate high number of travellers. Thus, the eleven bus stops will be kept in the model, but the risk of violating the assumption of residuals being normally distributed persists. A way to prevent violating the assumption is to look at the distribution of the variables. Variables with a skewed distribution, especially the dependent one, could disturb the normal distribution of the residuals. The extent to which a variable is skewed can be retrieved with SPSS and the values of skewness are displayed in 19. Preferably, skewness has a value between -1 and 1 (Lindner, 2013). In the appendix 19 it can be seen that rides exceed this, with a value of almost 18 and the skewness of

rides is visualized in figure 4. The figure shows the frequency distribution of the number of rides of the bus stops. The first bar shows how many times a value between zero and ten thousand rides is seen in the data. The y-axis is cut off at 50, but the bar continued to almost 900.

There are various ways of transformation that can reduce the skewness of variables such as the log-transformation or the Box-Cox transformation (Box & Cox, 1964). The log-transformation is easier to apply but the Box-Cox transformation, see appendix 9.4, was found for this study to reduce the skewness to a lower value than the log-transformation. Therefore, the Box-Cox transformation is applied to rides and to frequency, the latter because frequency is also a highly skewed variable and frequency will turn out later to be an important variable. Transforming all the highly skewed variables will make it harder to interpret the model, so the usage of transformation will be kept to a minimum. Applying the Box-Cox transformation is done in ArcGIS Pro and the result of applying it to rides is displayed below. The skewness is reduced to 0.013 (SPSS).

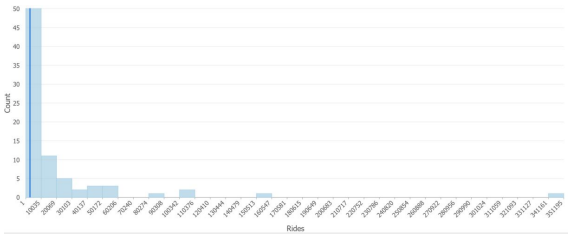

Figure 4: Histogram rides

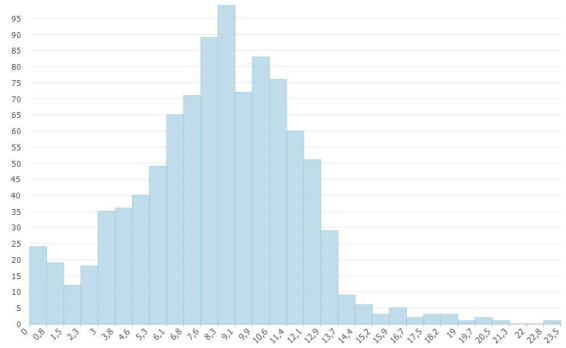

Figure 5: Histogram transformed rides

Furthermore, in this study Cook's distance is examined to identify multivariate outliers. Cook's distance is calculated per observation and shows how much the regression line is influenced by that observation (Cook, 1977). The calculation of Cook's distance includes all the independent values and the value for the dependent variable, in contradiction to a normal outlier for which only one observation for a single variable is examined. When an observation has a high value for Cook's distance it means that removing this observation will result in a large change in the regression model. Thus, Cook's distance can be used to identify missing variables. The following rule of thumb will be applied: if Cook's distance is higher than three times the mean of all Cook's distances it can be considered as a multivariate outlier (Boussiala, 2020).

3.4 Ordinary Least Squares

To see which variables affect bus ridership an OLS model will be estimated. To interpret the results of the OLS, the model should not violate any of the OLS assumptions. The OLS regression formula exists of four parts. Left of the equal to sign is the dependent variable (y). The size of this variable is explained by looking at independent variables (x). These variables have varying influences on y and these influences are represented by β, also known as coefficients. If a β is positive and the corresponding x increases, y will increase too. There are influences that affect y, but are not or could not be captured included in the model, these are represented in the error term ϵ. The α is a constant. The standard formula for OLS is displayed below

$$y = \alpha + \beta x + \epsilon \tag{1}$$

In this research multiple bus stops will be examined, so the number of y's will be the same as the number of bus stops in concession area Arnhem-Nijmegen (n). Looking at equation 1, this will make y an $n \times 1$ vector. X will be an $nx \times k$ matrix, where k is the number of independent variables. β will be a $k \times 1$ vector. ϵ will be in the same format as y, namely $n \times 1$.

3.4.1 Variable selection

Multiple OLS models will be estimated. The first model is used to identify potential new variables, by looking at patterns for example in the residuals or by looking at multivariate outliers. Then the new independent variables are added to the model and their influence is described. After reporting the influence of the variables on ridership per category, the variables with severe endogeneity problems are removed from the model. A full model is estimated and after the results are described, the OLS model is used to reduce the number of variables to make the model more clear and easier to use. To ensure the most influential variables are included and the loss of explanatory power is minimized, the variables are selected based on the 'forward selection' process in SPSS. This process starts estimating an OLS model with the variable that has the highest absolute correlation with the dependent variable (IBM, n.d.). SPSS displays the results of the model and it keeps adding variables one variable at the time. There are two conditions for adding a variable, the tolerance level should be above 0.0001 and adding the variable does not lead to already included values to drop their tolerance level below 0.0001. The choice can be made to include less variables than there are in the last step of the forward selection process, as the fewer variables are included in the model, the easier the model is to use. But, according to Prairie (1996), at least an R-squared of 65% is necessary for acceptable predictions and the prediction accuracy increases rapidly when the R-squared increases, will be kept in mind.

3.4.2 Significance

To conclude if there is actually a relationship between an independent variable and the dependent variable, the significance level is examined. The significance level is the probability of rejecting the null hypothesis, when the hypothesis is true. For every independent variable in this study, the null hypothesis is that there is not a relationship between the independent variable and the dependent variable. An example, for accessibility residents the null hypothesis is that there is not a relationship between accessibility residents and bus ridership. When this hypothesis is rejected it means that there is a relationship. If the significance level of accessibility residents of 0.005 means there is a 0.5% probability there is not in reality a relationship between accessibility residents and bus ridership while it is assumed there is. Thus, the lower significance the better. In this study a relationship is significant, when the significance level is below 0.1.

3.5 Spatial regression

In chapter 2 the existence of spatial auto-correlation was only assumed. The existence can be demonstrated by examining the residuals of the OLS model with the Moran I test (Griffith, 2000). The Moran I test, equation 2, examines neighbouring values of the residuals retrieved from the OLS, with the use of the spatial weight matrix W_{ij}, which will be explained below. In the numerator of the right fraction, it calculates when two stops are neighbours, how much for both neighbours, the values for the residuals differ from the mean. If both stops have either or considerably higher values or considerably lower values, the result will be a large positive value, which is thus an indicator the residuals of the neighbours show resemblance. The right denominator can be seen as a standardizer. The left part of the right side of the equation equals 1 in this study, so it can be neglected. To existence of spatial auto-correlation is proved when the Moran I test is significant.

$$I = \frac{N}{\sum_i \sum_j Wij} \frac{\sum_i \sum_j Wij(ei - \overline{x})(ej - \overline{x})}{\sum_i (xi - \overline{x})^2} \tag{2}$$

If the existence of spatial auto-correlation is proved, the spatial lag x and the spatial error model turned out to be the most suitable models for explaining bus ridership. The first model examines if the dependent variable is affected by the independent variables of a neighbour. The formula for this model is displayed below:

$$y = \alpha + \rho x Wij + \beta x + \epsilon \tag{3}$$

The βx has the same function as in the OLS model and can be labelled as the direct influence of the variable on the dependent variable. $\rho x W_{ij}$ can be described as the indirect influence, the influence that values for independent variables of neighbouring stops have on the ridership of a stop (Burkey, 2018). ρ, also called the spatial auto regressive parameter, is an estimation of the strength of the influence of neighbouring independent variables on the dependent variable (Gan et al., 2019). W_ij is the spatial weight matrix and will be explained below. The formulas to capture spatial auto-correlation in the error term, is presented below. Where the formula of the SLX looks like an OLS function with an extra factor, the SEM consist out of two formulas. This is because the SEM model bases the sizes of its coefficients on the spatial distribution of the error terms and in order to get an error term the observed value should be subtracted from a predicted value, therefore the first formula is needed.

$$y = \alpha + \beta x + \mu \tag{4}$$

$$\mu = \rho \mu Wij + \epsilon \tag{5}$$

The models will be estimated in Rstudio, based on a script of Burkey (2018).

3.5.1 Spatial weight matrix

The spatial weight matrix contains information if bus stops are neighbours of each other. It does not contain information on the extent to which neighbours are influencing each other, the rho is estimated to capture this. There are different kind of matrices, examples are one with the distances as the crows fly between stops or the travel times. Another matrix is filled with ones and zero's, ones if stops are directly connected to each other, zeroes if otherwise. Choosing the proper spatial weight matrix is of great importance, as the spatial weight matrix can have a substantial effect on the results (Tiefelsdorf, Griffith, & Boots, 1999). But support in choosing the right spatial weight matrix in literature is limited (Anselin, 2002). In this research, the neighbours of every stop are the three closest neighbours to the stop. The number of neighbours is deliberately set low, because of the spread of the stops as can be seen in 3. For the stops at the provincial roads a high number of stops would not be justifiable. In the spatial weight matrix, each row and each column represent a bus stop in Arnhem-Nijmegen, the length of the rows and the columns is thus 965. The matrix will be standardized for every neighbour each stop has. As the number of neighbours is fixed, the

matrix will mainly exists out of zeros but every row will have three 0.333 values if the stop in the column is one of the three closets stops.

3.6 Prediction

To know the predicted value for a bus stop the values for the independent variables, the corresponding coefficients and the constant α are filled in the equations described above. There are various metrics that can express the predictive capability of a model. The Akaike Information Criterion (AIC) can be used to compare the predictive power of different models on the same data set (Shmueli, 2010). Although the AIC does say something about the predictive power between models, the metric is not specific enough to know how a predictive model performs. Therefore, there will be looked at the root mean square error (RMSE) and mean absolute error (MAE). Each has its own focus on a certain aspect of the residuals (errors). For both goes, the lower the better.

MAE is a metric for the average size of the residuals. It sums the residuals and divides it by the number of observations (Chai & Draxler, 2014). To calculate the RMSE, the residuals are first squared, before they are divided by the number of observations. This way larger residuals are penalized. For policymakers an overestimate of 200 rides could be worse than two overestimates of 100, therefore will be looked at the RMSE. But a high RMSE could be caused by one exceptional case. That is why both metrics are useful for predicting bus ridership at stop level.

The RMSE and MAE are suitable metrics to express the accuracy of the model, but these single values are not useful to study the prediction accuracy. When the residual and the relative residual are examined, it is possible to see in which areas the model performs well and in which area it performs less well. Both are necessary to study, a residual of 500 may not seem like a lot spread over the 21 week days in November but if the actual ridership was 5 it can be seen as a serious mismatch. On the other hand, in the data there are some stops with 1 ride for the whole month November, if the model predicts 20 rides it results in a relative residual of 2000%.

The relative residual is retrieved by subtracting the observed value from the predicted value and then the retrieved value is divided by the observed value, times 100%.

Rstudio will be used to estimate the three regression models, to retrieve the metrics and to retrieve the prediction intervals.

3.7 Conclusion

For deciding which variables to include in the model, the availability of data is of a major importance. The fare tickets of the bus are the same in the study area, therefore it is redundant to include this variable. This all resulted in a selection that can be categorised in three groups, demographics, built environment and level of service. In table 8 the variables are displayed, which will be included in the models but were not found the in literature review.

Table 8: Included variables, which were not identified in the literature review

Demographics	Built environment	Level of service
Green votes	LU amusement	Bicycle stand
	Secondary education within buffer	Accessible for disabled
	Closest to train station	
	Closest to secondary education	
	Closest to higher education	
	Closest to secondary education	
	Closest to P+R	
	Hospitals accessible	
	Higher education accessible	
	Secondary education accessible	
	P+R accessible	

Statistical models will be used with two purposes in this study, explaining and predicting. For this first purpose, the assumptions of OLS should not be violated in order to get the best estimates for the coefficients. The first part of the next chapter discusses the OLS results of the first model, which is based on the variables mentioned in this chapter. This model will be evaluated and checked where the model performs poorly. New variables that could improve the performance are discussed and added the model. Then the relationships between the independent variables and dependent variables are discussed per variable category. This way, insights are retrieved which variables per category are affecting ridership the most. After some variables have been disregarded, three full models (OLS, SLX and SEM) and their results are interpreted. To make the prediction model easy to use, the number of explanatory variables is reduced and new OLS and SLX models will be estimated. The process of the different models and the flow of variables is summarized in 6.

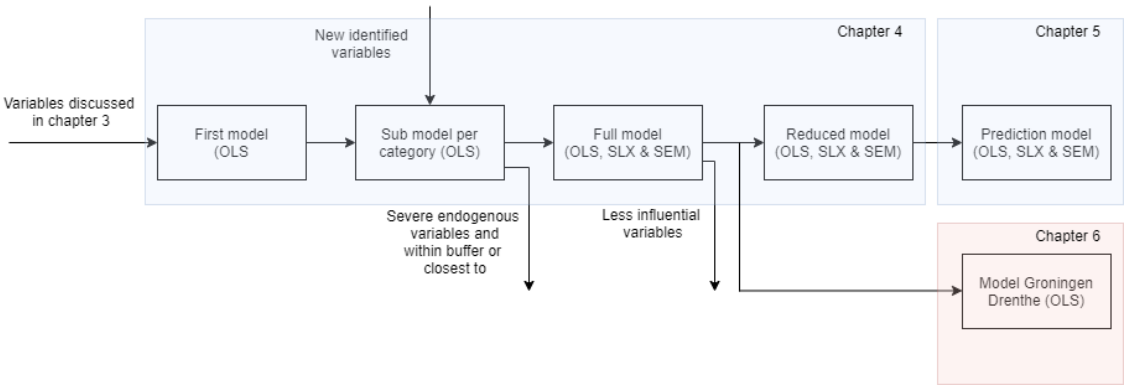

Figure 6: Conceptualisation models

4 Explaining bus ridership

In this chapter multiple regression models will be estimated. The first model estimated, will not be discussed in depth, because it will mainly be used to examine if there are other explanatory variables missing. Afterwards, a sub-model is estimated for each of the variable categories to have a clear view on the relationships between the explanatory variables and bus ridership. Then the full model with all the variables included, is examined. Comparisons are made between the significance of the relationships in the sub-models and in the full model and the extent of which the independent variables affect the ridership. This, to get a better understanding how the model behaves. Subsequently, the need for spatial regression models is proved, after which spatial models are estimated. The results are interpret and compared with the OLS model.

4.1 Results first model

The first model has a R-squared of 74.41%, which means that 74.41% of observed variation can be explained by the explanatory variables and the values for the coefficients and the significance levels are displayed in the appendix 21. To discover missing explanatory variables, the first step is examining the standardized residuals in ArcGIS Pro to see if there are any patterns. In figure 7, the bus stops are displayed and a red color stands for that the estimated number of rides is below the actual number and blue represents an overestimate of the number of rides.

What is striking is the number of blue dots in Dieren, east of National Park de Veluwezoom and the blue dots in the direction to Doetinchem. A reason could be road constructions, but according to https://wiki.ovinnederland.nl/ the buses were driving their normal schedules around Dieren and Doetinchem. Another reason could be the multiple bus operators stopping at the same stop, for the road to Doetinchem it is Breng and Arriva and for Dieren it is Breng and Syntus. The smart card data set only contains the rides for Breng. Thus, the *number of bus operators* that stop at a bus stop could be a valuable explanatory variable and this variable has not been studied before in other studies. Other blue dots that caught the eye were the stops above the river, north of Nijmegen. This led to the idea to examine the *difference in travel time between bus and bike*, something which has not been examined before as well. Furthermore, the model seems to underestimate the number of *rides* in the city centres of Arnhem and mostly in Nijmegen. Apparently, *address density*, *number of jobs*, *social and commercial land-use* and *frequencies* are not enough to capture the city center effect. The variable *distance to city center*, seen in chapter 2, could have been helpful for this.

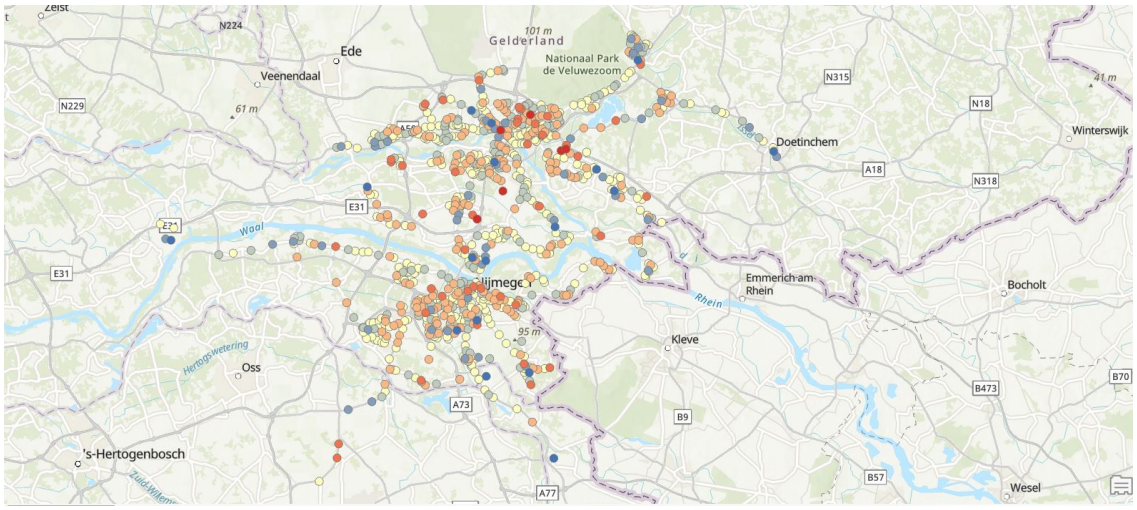

Figure 7: Standardized residuals

When examining the standardized coefficients and the significance levels of the explanatory variables in appendix 21, it is noticeable that there is a positive significant relationship between *ridership* and if a stop is *accessible to disabled*. In section 1.1.1, we saw that the disabled travel less and the public transport does not play a bigger part for them. Thus, it is likely that the positive relationship between stops being accessible to disabled has a different cause. Looking at figure 8 it is hard to find a pattern, the red dots mean that the stop is *accessible for disabled*. Many of the stops in Arnhem and Nijmegen are accessible for disabled and have a high ridership, but there are other independent variables, such as *address density* or the *number of jobs*, that should be able to explain the high ridership. Furthermore, there seems a cluster of non-accessible stops for disabled in Dieren, but the overestimating is likely to be caused by the multiple *bus operators* at a stop. The provincial roads seem to have vary in *accessibility for disabled*, which makes it hard to find a new explanatory variable.

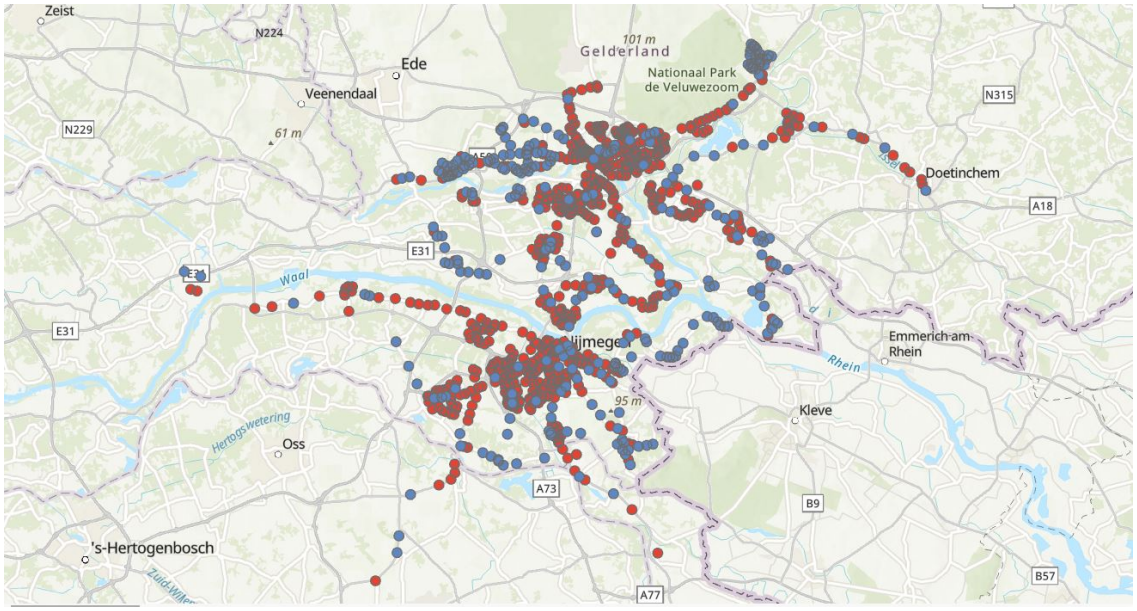

Figure 8: Stops that are accessible for disabled

The last way new to identify new variables, is by looking at Cook's distance. In figure 9 all the stops with a Cook's distance three times greater than the mean are displayed in red. By closer inspection in ArcGIS Pro it appeared that the stops at Arnhem central station and Nijmegen central station are multivariate outliers. The bus stops at these stations have without a doubt the highest number of *rides*, namely 350,000 and 150,000. For comparison, the stop with the tenth highest number of *rides*, has 47,000 *rides*. Apparently the *closest to station*, *address density* and *frequency* variables are not enough to explain the high number of rides at these stops. A reason for this, can be found in the number of train passengers of the stations in the concession area. The closest bus stop to train station Arnhem Centraal is likely to cause a higher number of bus passengers than the train station in Dieren. Thus, this should be taken into account in the model.

The stop with the second highest Cook's distance, the highest was Nijmegen Centraal, was a stop Venlo in the south of the Netherlands. Venlo is included in the smart card data, because of a bus route 83 going from Nijmegen to Venlo and vice versa. The data contains only one boarder in Venlo for the month November, something that seems highly irregular for a stop right in the heart of Venlo, with a population over 100,000 (Alle Cijer, 2021). The bus route is shared by Arriva and Breng, which could explain the discrepancy. Because of the unlikeliness of only one boarder and the influence the stop has on the regression model, the choice has been made to exclude the stop.

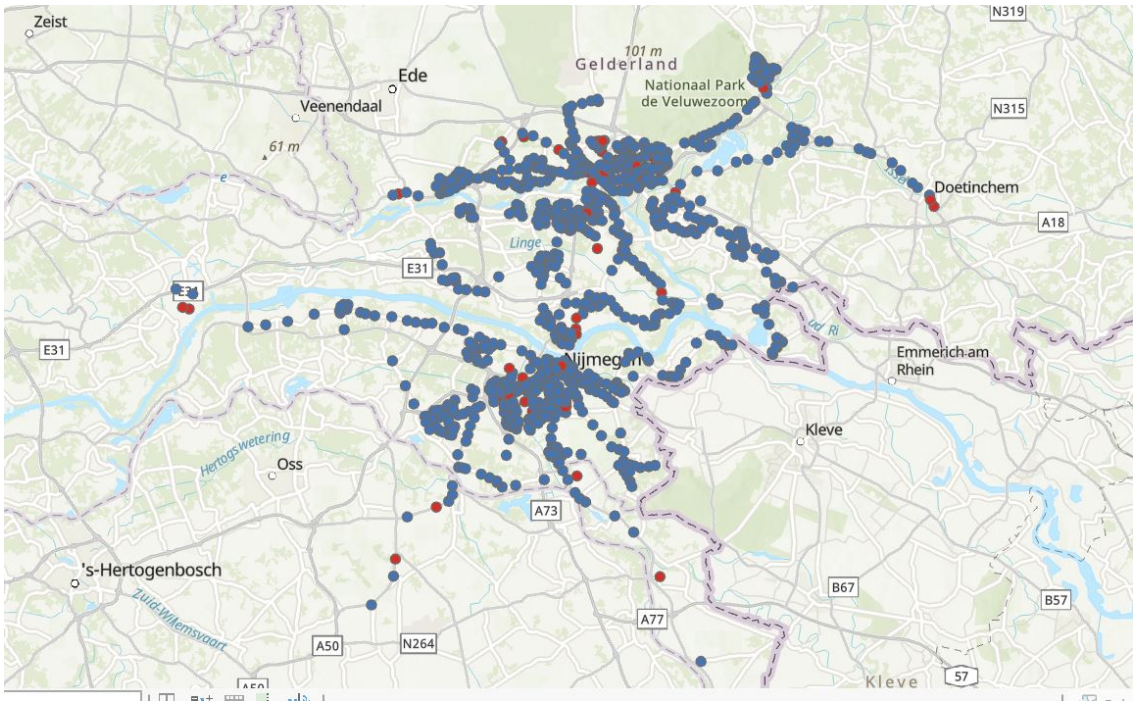

Figure 9: Cook's distance

4.1.1 Implementing new variables

For data on the *number of bus operators* stopping at a bus stop the same data set used as for *frequency*, namely the GTFS data set (Open mobility data, 2019). It does not contain directly the *number of operators* and Rstudio is used to retrieve this information. The variable turned out to be significant, but unfortunately it did lead only to an increase of 0.36% in the R-squared to 74.7%. A reason for this, could be that only 192 out of the 965 bus stops has multiple bus operators.

To see what the effect is of the *difference between bike time and bus time* on *bus ridership*, for each stop the travel time to the nearest train station is calculated with the use of the VBW. Then the travel time for the bike is subtracted from the travel time of the bus. This difference is added as variable to the model and it turned out that the bike is the faster mode for 23% of the bus stops. But as can be seen in figure 22, the number of rides of bus stops where the bike is faster is broadly spread and it is impossible to find a linear relationship. The relationship between *difference in travel time* and *bus ridership* turned out to be insignificant.

To deal with the high number of bus passengers at large train stations, the interaction term *closest to station x address density* is added, since data on the train passengers per train station was not available for the train operator Arriva. This interaction term turned out to be insignificant. Including a dummy variable if a stop was at a *central station* was more successful. The R-squared increased to 0.751.

4.2 Regression model per category

The next models contain subsets of the explanatory variables described in chapter 3, without *residents* and *accessibility residents* and is complemented with the *number of bus operators* at a stop and a dummy if the stop is at a *central station*. In this section a more in depth explanation of the model will be given and afterwards the full model is discussed.

4.2.1 Demographics

First the influence of demographics on *bus ridership* is analysed with the use of SPSS. It turned out that demographics can explain 26% of the *bus ridership.* In the table below, it can be seen that two variables have a significance level under 0.01, namely *average car ownership* and the *percentage of elderly*. The negative relationship between *elderly* and *ridership*, is logical as described in section 1.1.1. And if the *average car ownership* is low, the *bus ridership* is high, which is logical as well. The *percentage female* and *percentage of residents with a Western background*, excluding Dutch, are both positively affecting *ridership*, with a significance level under 0.1. The other independent variables are not significant. Thus, this means this model cannot say with high certainty that there is a relationship between *ridership* and *percentage youth*, *percentage green votes*, *percentage of residents with a non-Western background* and *average household size*. To determine which of the significant variables is affecting *ridership* the most, the absolute value of the standardized coefficients (Beta) is examined. *Average car ownership*, turned out to affect *ridership* the most, then *percentage elderly*, *percentage female* and *percentage Western background.*

Table 9: Demographics

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	5.254	3.839		1.368	0.171
Average household size	0.101	0.475	0.011	0.213	0.832
Average car ownership	-4.139	0.977	-0.300	-4.237	0.000***
Female	0.104	0.054	0.066	1.930	0.054*
Youth (15-24)	0.027	0.025	0.043	1.084	0.279
Elderly (65+)	-0.066	0.017	-0.149	-3.977	0.000***
Households with low income	0.023	0.015	0.077	1.542	0.123
Residents Western background	0.086	0.050	0.055	1.718	0.086*
Residents Non Western background	-6.206E-05	0.015	0.000	-0.004	0.997
Green votes	0.014	0.012	0.050	1.207	0.228

Significance levels: * 0.05 - 0.1, ** 0.01 - 0.05, *** <0.01

4.2.2 Built environment

The R-squared, or in other words the explanatory power of the built environment variables is 43.7%. From the 27 variables, thirteen variables have a significance level below 0.05 and two have a significance level below 0.1. *Residential land-use* appeared to affect *ridership* the most and thereafter s*ocial and commercial land-use*. In agricultural regions the bus ridership is low, as expected. *Recreational land-use* (sport facilities and parks) on the other hand seem to have, surprisingly, a positive influence on *ridership*. Furthermore, the higher the *address density* or the *parking cost*s, the higher the *ridership*. There are two causes for the relationship between *parking costs* and *ridership*. The first one is because of the high *parking costs*, people choose the bus over the car or the *parking costs* are high because the stop is in a popular area, which explains the higher *ridership*. *Distance to station* turned out to have a significant effect, in contrary with the findings of Guo and Huang (2020).

The *within buffer* variables turned to be better at explaining ridership than the *close to* variables. Of the latter group, only *closest to train station* turned out to be significant, while for the *within buffer* group train stations, hospitals and higher education are significant. It is noticeable that train station appears in both and that their VIF factors are low. Examining the *frequencies* of those variables, shows that there are 23 stops that are the closest stop for a train station and 52 stops that have a train station within their catchment area. It turns out that for a stop, being close to a station, but not being the closest has a negative influence on the ridership. This seems sensible, as many travellers will not take the bus for one stop, when heading to the station. The absolute value for standardized coefficient for *closest to* is higher than the one for *within buffer*, so the effect of a train station for the closest bus stop is still positive on bus ridership, especially

for the central stations. The presence of *competitive bus stop in the catchment area* turns out to have a negative effect on the number of *rides*.

Applying the Forward Selection confirms above described. For the facilities, except for the train station, the within buffer will be included in the full model. For the train station closest to will be included.

The amount of *jobs can be reached from a bus stop*, is the most influential built environment variable. The higher the amount of *jobs that can be reached*, the higher number of *bus rides*. The amount of *higher education facilities a traveller can reach* from a stop affects *bus ridership* positively as well.

Table 10: Built environment

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Constant)	4.406	0.270		16.304	0.000***
LU residential	0.013	0.004	0.101	2.882	0.004***
LU business	0.006	0.010	0.014	0.555	0.579
LU social and commercial	0.034	0.019	0.058	1.807	0.071*
LU agricultural	-0.018	0.008	-0.053	-2.099	0.036**
LU recreational	0.022	0.011	0.052	1.954	0.051*
LU amusement	0.026	0.028	0.023	0.914	0.361
Address density	4.781E-05	0.000	0.087	2.191	0.029**
Distance to station	0.101	0.017	0.178	5.800	0.000***
Parking rate	75.340	30.513	0.064	2.469	0.014**
Jobs within buffer	1.985E-05	0.000	0.006	0.172	0.863
Train stations within buffer	-1.139	0.555	-0.073	-2.052	0.040**
Hospitals within buffer	2.243	0.931	0.076	2.408	0.016**
Higher education within buffer	2.950	1.089	0.089	2.708	0.007***
Secondary education within buffer	0.377	0.251	0.049	1.503	0.133
P+R within buffer	0.034	0.741	0.002	0.046	0.963
Other bus stops within buffer	-0.396	0.086	-0.126	-4.600	0.000***
Closest to train station	3.945	0.803	0.171	4.910	0.000***
Closest to hospital	1.532	1.512	0.031	1.013	0.311
Closest to higher education	-1.788	1.788	-0.033	-1.000	0.318
Closest to secondary education	0.021	0.395	0.002	0.053	0.958
Closest to P+R	0.824	1.185	0.022	0.695	0.487
Jobs accessible	4.787E-05	0.000	0.324	6.760	0.000***
Hospitals accessible	-0.162	0.154	-0.037	-1.051	0.294
Higher education accessible	0.420	0.144	0.099	2.905	0.004***
Secondary education accessible	0.033	0.032	0.066	1.019	0.308
P+R accessible	-0.028	0.142	-0.011	-0.197	0.844
Dummy central station	5.698	2.164	0.074	2.632	0.009***

Significance levels: * 0.05 - 0.1, ** 0.01 - 0.05, *** <0.01

4.2.3 Level of service

The category with the highest R squared (69.6%) is level of service, and that is mostly due to the *frequency*. *Frequency* has by far the highest standardized coefficient and is positively related to *ridership*. The *number of directions* is positively related too. When only the smart card data of one agency is examined, *the number of operators* has a negative effect on ridership, just as expected. When the smart card data of all the bus operators is included in a model, the *number of operators* is expected to have a minor or even a positive influence of the number of *rides*.

As mentioned in chapter 3, the endogeneity problem arises the most with the level of service variables. Chances are high that a *digital travel information display* is placed, because a bus stop has many *rides*, just as with *frequency*, *number of directions*, *shelter* and *benches*. This does not have to apply to *bicycle parking* and *accessible for disabled*. But the first one is not significant and for the second it is doubtful the effect is truly caused by bus stops being *accessible for disabled*. The objective of this study is to predict the number of rides and because of the endogeneity problem the physical bus stop characteristics are not suitable for prediction. When a new stop is planned, the choice of whether to add a *digital information display* to the bus stop or not would affect the expected number of rides. The values for the variables *frequency*, *number of directions* and *number of operators* can be more substantiated, since new bus stops should be integrated in existing bus routes or new bus routes should be designed. Therefore, these three variables are still included in the model.

Table 11: Level of service

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	4.592	0.214		21.443	0.000***
Frequency	1.970	0.088	0.584	22.484	0.000***
Directions	0.081	0.031	0.067	2.641	0.008***
Bus operators	-0.800	0.162	-0.106	-4.947	0.000***
Shelter	0.215	0.223	0.029	0.968	0.333
Bench	0.974	0.227	0.131	4.286	0.000***
Digital information display	1.520	0.168	0.188	9.044	0.000***
Bicycle stand	-0.121	0.153	-0.015	-0.793	0.428
Accessible for disabled	0.892	0.162	0.116	5.496	0.000***

Significance levels: * 0.05 - 0.1, ** 0.01 - 0.05, *** <0.01

4.3 Full model OLS

The full model exists out of the variables discussed in the previous three models. First it will be examined if the assumptions for multiple regression are violated. Thereafter, the results of the full model will be discussed.

4.3.1 Multiple regression assumptions

One of the multiple regression assumptions is that there is no multicollinearity in the data. In table 12 it can be seen that no variable has a VIF value over ten and only three variables have a VIF higher than five. There are two plots in the appendix to check if the residuals of the OLS model are normally distributed and if the variance of the residuals is constant. The conclusion can be drawn, that there is no severe violation of the two assumptions. Moreover, the values of the residuals should be independent and to check this the Durbin Watson value is used. The Durbin Watson value can take values between zero and four and two is the most optimum value. This model has a Durbin Watson of 2.05 see Appendix. But this does not mean the assumption is met, because the Durbin-Watson is unable to capture spatial auto-correlation in the error terms. After discussing the results of the full OLS model, it is examined if there exists spatial auto-correlation in the error terms.

4.3.2 Results

A model including all the demographics, built environment and level of service variables has an explanatory power of 75.8%, which is 1.4% lower than the explanatory power of Kerkman et al. (2015). One of the reasons for the discrepancy is not including the *distance to center* variable, which was found to have a large influence on bus ridership in the study of Kerkman et al..

After removing the physical stop characteristic, such as the presence of a *shelter*, there are 34 variables left which can explain 71.4% of the ridership. From the nine demographics variables, four are significant, see table 12. It is shown that a higher *car ownership* will results in a lower bus *ridership*. In section 1.1.1, we saw that the percentage of trips made by bus for *elderly* was similar to the average Dutch person. The negative relationship between *elderly* and bus *ridership* can be explained by that the *elderly* travel less. *Ethnicity* and *percentage of households with a low income* do not influence *bus ridership* significantly, in contrast to findings in chapter 2. Apparently, there are no cultural differences in the Netherlands when it comes to bus travel behaviour. Kerkman et al. (2015) did find a significant negative relationship between *income* and *bus ridership*, but in their model *car ownership* was not included. In appendix 9.11, it can be seen that *percentage low income households* and *car ownership* are strongly correlated and in this study *car ownership* is more important for *bus ridership* than *percentage low income households*.

The land-use variables are performing well, only *business* and *agricultural* are not significant. The definition for business is probably too large, as the data includes offices as well as industrial companies. Two land-uses that could have different kinds of bus travel behaviour. *Agricultural* was expected to have a negative influence on the number of *rides*, but this relationship is not significant.

Address density, the *parking rate* and *jobs within buffer* have lost their significance. This is likely due to *frequency* being included in the model, as the *address density*, *parking rate* and the *number of jobs* are high in areas where the *frequency* of the busses is high as well. What the model looks like without *frequency* is displayed in table 22 in the appendix and it can be seen that *address density*, *parking rate* and *jobs within buffer* are significant.

The results for *higher education within buffer* and secondary education within buffer are aligned with the numbers of CBS, seen in section 1.1.1. The presence of these educational facilities does increase the ridership. It is interesting to see that *Park and Ride* is the only facility without a significant relationship. Two reasons for this could be that the *Park and Rides* in Arnhem-Nijmegen are not attracting many travellers or the travellers that do park at the *Park and Rides* choose to not take the bus.

A model with *frequency* and *accessibility higher education*, resulted in the loss of significance for the latter. This is likely to be caused by the fact that bus stops on the routes to higher education facilities have a high *frequency* and thus including *frequency* in the model makes the *accessibility of higher education redundant*. The relationship between *rides* and *accessibility of hospitals* stayed significant. In the appendix 9.12 it is visible that the correlation between *frequency* and *accessibility higher education* is higher than between *frequency* and *accessibility hospitals*.

Table 12: Full model

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	Sig.	VIF
(Constant)	2.283	2.636		0.387	
Average household size	0.610	0.315	0.062	0.053*	3.378
Average car ownership	-1.346	0.687	-0.095	0.050**	7.672
Female	0.018	0.038	0.011	0.634	1.822
Youth (15-24)	0.039	0.018	0.06	0.035**	2.583
Elderly (65+)	-0.032	0.012	-0.07	0.007***	2.190
Households with low income	0.009	0.010	0.029	0.404	3.770
Western background	0.036	0.034	0.023	0.288	1.466
Non Western background	-0.001	0.010	-0.002	0.946	2.769
Green votes	0.004	0.008	0.013	0.652	2.745
LU residential	0.021	0.004	0.165	0.000***	2.548
LU business	-0.011	0.008	-0.026	0.184	1.201
LU social and commercial	0.028	0.014	0.047	0.049**	1.841
LU agricultural	-0.005	0.006	-0.016	0.396	1.096
LU recreational	0.018	0.008	0.041	0.032**	1.210
LU amusement	0.039	0.021	0.035	0.063*	1.124
Address density	-1.332E-05	0.000	-0.025	0.470	4.003
Distance to station	0.045	0.013	0.078	0.001***	1.731
Parking rate	11.374	22.532	0.009	0.614	1.122
Jobs within buffer	1.559E-05	9E-5	-0.024	0.368	2.223
Hospitals within buffer	1.815	0.558	0.06	0.001***	1.112
Higher education within buffer	1.979	0.678	0.058	0.004***	1.291
Secondary education within buffer	0.281	0.148	0.036	0.058*	1.159
P+R within buffer	-0.208	0.442	-0.009	0.638	1.228
Other bus stops within buffer	-0.109	0.065	-0.034	0.095*	1.341
Closest to train station	1.145	0.464	0.049	0.014**	1.252
Jobs accessible	0.000	0.000	0.093	0.010***	4.236
Hospitals accessible	-0.218	0.116	-0.049	0.061*	2.175
Higher education accessible	0.095	0.112	0.022	0.396	2.181
Secondary education accessible	-0.012	0.024	-0.024	0.602	7.078
P+R accessible	-0.098	0.108	-0.037	0.365	5.460
Dummy central station	2.882	1.762	0.036	0.102	1.603
Frequency	2.586	0.106	0.632	0.000***	2.186
Directions	0.064	0.038	0.052	0.090*	3.052
Bus operators	-0.447	0.175	-0.058	0.011**	1.662

Significance levels: * 0.05 - 0.1, ** 0.01 - 0.05, *** <0.01

4.3.3 Conclusion

So far, new explanatory variables have been identified, *dummy central station* and the *number of operators*. Several sub models have been estimated to retrieve an understanding how the model behaves. It turned out that the *within buffer* variables outperformed the newly examined *closest to* variables, only for train stations pointing out the closest stop was an improvement over the within buffer variant. Furthermore, the level of service variables were found to be able to explain *bus ridership* the most. But the endogeneity problem should be kept in mind.

For the full model the severe endogenous bus characteristics were disregarded, because it is hard for these characteristics to substantiate the choice of adding them when a new bus stop is planned. *Frequency* turned out to be all-encompassing, which is rather logical as the *frequency* is determined based on previous ridership. When the previous ridership is high, the *frequency* is likely to be increased.

For the variables *address density*, *parking rate*, *jobs within buffer* and *accessibility higher education* a significant relationship with *bus ridership* is dependent on whether *frequency* is included. Newly identified significant relationships are *land-use amusement secondary education within buffer* and *accessibility of hospitals*, which are all positively related to *ridership*. For political preference expressed in the *percentage green votes*, the presence of *bicycle stand*, *secondary education accessible* and *P+R accessible* no significant relationship with rides was determined.

The assumptions for multiple regression have been discussed, but spatial auto-correlation in the error term could lead to false interpolation of the results. Therefore in the next section, the spatial models are discussed.

4.4 Full model SLX & SEM

In chapter 2 the existence of spatial auto-correlation was only assumed. The existence can be demonstrated by examining the residuals of the OLS model with the Moran I test, see section 4.4. Significance of the Moran I test means that the hypothesis that there is not spatial auto-correlation in the error terms can be rejected. The test was conducted in Rstudio and the p-value was 3.2e-11, thus it can with a high certainty be concluded there exists spatial auto-correlation and spatial models are required.

The same independent variables as seen in the OLS full model are used for the SLX and SEM models. For the SLX model, the list of independent variables is completed by adding the neighbouring values of those independent variables. The OLS and SEM model do not take those neighbouring values into account, therefore it is not possible to compare the estimates of the OLS and SEM models with the direct effects of the SLX model. For that reason, the OLS and SEM models will be compared first and afterwards the results of the SLX model are discussed.

4.4.1 Spatial Error Model

The SEM model has a explanatory power of 72.4% and the estimates are shown in table 13. It is striking to see that some variables are significant in the OLS model and are not significant in the SEM model. The technical reason for this is that the clustering of the residuals has led to different estimates and different significance levels.

One of the variables that is not significant in the SEM model is *secondary education within buffer* and this variable will be used as example to demonstrate how a difference in significance level can be caused. When the spatial distribution of bus ridership, figure 3, and the locations of *secondary schools* are compared, figure 25 in appendix 9.13, the schools are often found where the ridership is high. The OLS model, which does not take locations into account, concludes that a secondary school in the catchment area has a positive effect on the number of rides. But the SEM model looks at the residuals of the neighbours, which could have no *secondary schools* in the catchment area, and sees that the model underestimates the number of rides for these stops. Therefore, the model can conclude that the high number of rides is not caused by the presence of the *secondary school*, but it is caused by a different variable.

The same principle goes for *car ownership*. In appendix 9.14 it is visible that *car ownership* is high in the smaller towns and low in the cities, which is the opposite pattern of *rides*. The OLS model concludes that there is a negative relationship between rides and *car ownership*. The SEM model sees that the bus stops with high *car ownership* are spatially related and concludes the lower number of rides in villages is not caused by the high *car ownership*.

The reason for the loss of significance for the accessibility variables, could be that the limit on access time to reach the bus stop was set too large in the VBW. In ten minutes it could be possible to reach another stop. Accessibility values for neighbouring stops could therefore be very similar and this is probably the reason why the SEM model concludes that the relationships between *rides* and the accessibility variables are not significant. A more elaborate explanation is provided in appendix 9.15.

Table 13: OLS and SEM Full model

	OLS			SEM		
	B	Std. Error	Sig.	B	Std. Error	Sig.
(Constant)	2.283	2.636	0.387	1.562	2.878	0.587
Average household size	0.610	0.315	0.053*	0.639	0.344	0.063*
Average car ownership	-1.346	0.687	0.050**	-0.930	0.737	0.207
Female	0.018	0.038	0.634	0.009	0.041	0.830
Youth (15-24)	0.039	0.018	0.035**	0.053	0.021	0.012**
Elderly (65+)	-0.032	0.012	0.007***	-0.024	0.013	0.073*
Households with low income	0.009	0.010	0.404	0.011	0.012	0.356
Western background	0.036	0.034	0.288	0.029	0.037	0.431
Non Western background	-0.001	0.010	0.946	0.004	0.012	0.765
Green votes	0.004	0.008	0.652	0.007	0.010	0.450
LU residential	0.021	0.004	0.000***	0.027	0.004	0.000***
LU business	-0.011	0.008	0.184	-0.009	0.008	0.297
LU social and commercial	0.028	0.014	0.049**	0.042	0.015	0.005***
LU agricultural	-0.005	0.006	0.396	-0.003	0.007	0.649
LU recreational	0.018	0.008	0.032**	0.019	0.009	0.029**
LU amusement	0.039	0.021	0.063*	0.039	0.023	0.094*
Address density	-1.332E-05	0.000	0.470	2.325E-5	2.300-05	0.312
Distance to station	0.045	0.013	0.001***	0.039	0.016	0.014**
Parking rate	11.374	22.532	0.614	14.073	20.894	0.501
Jobs within buffer	1.559E-05	9E-5	0.368	0.000	0.000	0.395
Hospitals within buffer	1.815	0.558	0.001***	1.796	0.600	0.003***
Higher education within buffer	1.979	0.678	0.004***	1.862	0.737	0.012**
Secondary education within buffer	0.281	0.148	0.058*	0.216	0.156	0.165
P+R within buffer	-0.208	0.442	0.638	-0.273	0.472	0.563
Other bus stops within buffer	-0.109	0.065	0.095*	-0.051	0.067	0.445
Closest to train station	1.145	0.464	0.014**	1.257	0.433	0.004***
Jobs accessible	1.401E-5	5.464E-6	0.010***	9.502E-6	6.307E-6	0.132
Hospitals accessible	-0.218	0.116	0.061**	-0.214	0.134	0.111
Higher education accessible	0.095	0.112	0.396	0.073	0.128	0.570
Secondary education accessible	-0.012	0.024	0.602	-0.013	0.028	0.642
P+R accessible	-0.098	0.108	0.365	-0.108	0.123	0.380
Dummy central station	2.882	1.762	0.102	1.462	1.617	0.366
Frequency	2.586	0.106	0.000***	2.654	0.109	0.000***
Directions	0.064	0.038	0.090*	0.094	0.038	0.012**
Bus operators	-0.447	0.175	0.011**	-0.530	0.184	0.004***

Significance levels: * 0.05 - 0.1, ** 0.01 - 0.05, *** <0.01

4.4.2 Spatial Lag X model

The number of independent variables has doubled in the SLX model. The extra added variables are the variables for the neighbouring values, the indirect effects. To reduce multicollinearity, nine neighbouring independent variables have been excluded from the model. The SLX model can be written as an OLS model, which makes it possible to conduct a Moran I test for the SLX model. The significance level of the test is very low, namely 5.96e-6. This means that the probability of spatial auto-correlation in the error term is still high and that the SLX model does not include all the explanatory variables that cause the spatial auto-correlation.

The SLX model has the highest explanatory power of the three models (73.4%) and the estimates of the SLX model are shown in appendix 9.16. Interesting to see, but not unexpected, is the significant negative relationship between rides and the number of directions of neighbouring bus stops. This means that travellers are willing to go to a different stop, if this stop has a higher number of directions.

More surprising are the significant negative relationships between *ridership* and *neighbouring residential land-use* and *ridership* and *neighbouring social and commercial land-use*. When a SLX model is estimated without *neighbouring residential land-use* the coefficient *residential land-use* is estimated lower than when *neighbouring residential land-use* is included. Apparently, the model gets better results by assigning larger coefficients to *residential land-use* and compensating this by *neighbouring residential land-use*. The same goes for *social and commercial land-use*.

4.5 Reduced model

A prediction model with 34 explanatory variables is hard to use and many variables will have a limited influence on the prediction accuracy. To reduce the number of variables, the forward selection process is applied. The values for the R-squared and the AIC for estimated models in the process are displayed in figure 10. After including six variables the increasing of the R-squared and the decreasing of the AIC seem to have slowed down significantly. To improve the prediction power, the choice has been made to include twelve variables in the reduced model, as the R-squared is not remarkably high. Removing the 23 variables resulted in a decrease in explanatory power of 1.2% to 70.2% for the OLS model and a decrease of 1.5% to 71.9% for the SLX model.

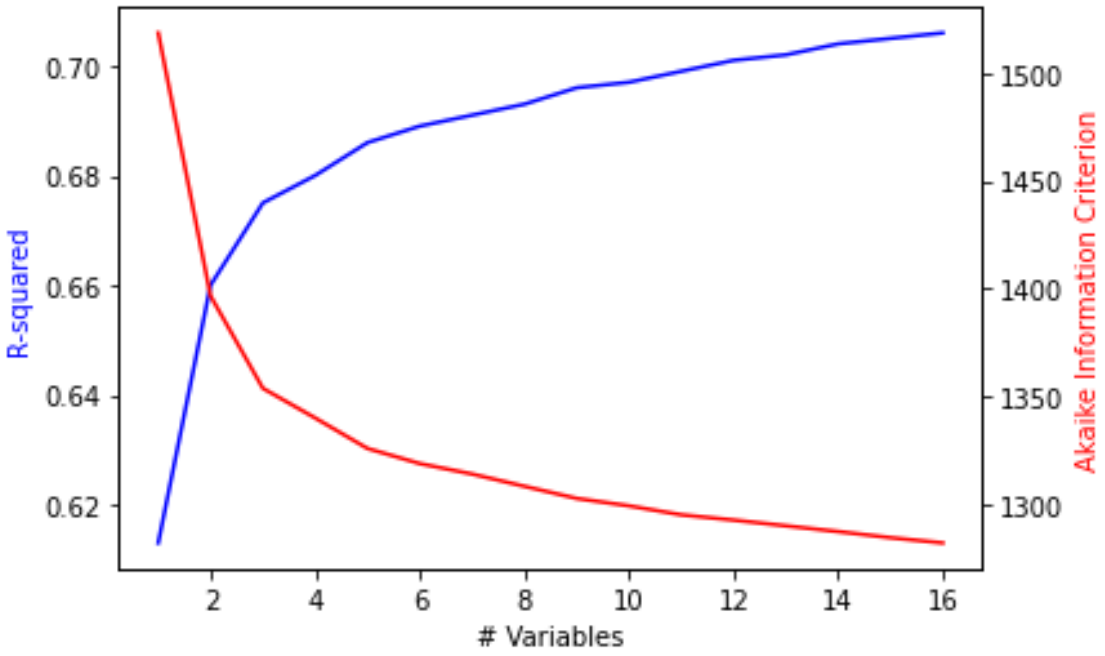

Figure 10: Forward selection process

The twelve variables and their estimates of the OLS and SLX models are displayed in table 14. The SEM model is disregarded as this model will not be used for predictions. It is noticeable that land-use for amusement parks and zoos was significant in the full model, but is not included in the reduced model. The reason for this is probably the fact that the area Arnhem-Nijmegen only consists of a small part of amusement. The same goes for the dummy for central station, which only concerned two out of 964 stops. Apparently, the added value of youth and average household size is too limited to be included as well. For the SLX model it is noticeable that Lag x Frequency is significant, while it was not significant in the full model. The reason for this is probably the exclusion of the directions variable and thereby the exclusion of neighbouring directions.

Table 14: Estimation results reduced model OLS and SLX

	OLS		SLX	
	B	Sig.	B	Sig.
(constant)	5.772	0.000***	6.496	0.000***
Average car ownership	-1.390	0.000***	-0.720	0.316
Elderly (65+)	-0.034	0.000***	-0.012	0.421
LU residential	0.019	0.000***	0.034	0.000***
LU recreational	0.016	0.040**	0.014	0.179
Distance to station	0.049	0.000***	0.028	0.273
Hospitals within buffer	1.975	0.000***	1.476	0.048**
Higher education within buffer	2.717	0.000***	1.936	0.034**
Closest to train station	1.563	0.000***	1.583	0.000***
Jobs accessible	0.000	0.003***	0.000	0.036**
Hospitals accessible	-0.278	0.004***	-0.140	0.472
Frequency	2.712	0.000***	2.821	0.000***
Operators	-0.279	0.064*	-0.133	0.457
Lag x Average car ownership			-0.738	0.366
Lag x Elderly (65+)			-0.024	0.181
Lag x LU residential			-0.026	0.000***
Lag x LU recreational			0.003	0.825
Lag x Distance to station			0.031	0.232
Lag x Hospitals within buffer			1.312	0.204
Lag x Higher education within buffer			1.014	0.360
Lag x Closest to train station			-0.028	0.972
Lag x Jobs accessible			0.000	0.000***
Lag x Hospitals accessible			-0.090	0.677
Lag x Frequency			-0.405	0.005***
Lag x Operators			-0.400	0.099**

Significance levels: * 0.05 - 0.1, ** 0.01 - 0.05, *** <0.01

4.6 Conclusion

Three models (OLS, SLX and SEM) have been estimated to explain bus ridership. Including spatial relationships turned out to improve the explanatory power with 1% (SEM) and 2% (SLX) compared to the OLS model. But more importantly, some of the independent variables, such as *car ownership* and *secondary education within buffer*, were found to be significantly related to ridership in the OLS model turned out to be not significant in the SEM model. Because the SEM model takes the residuals of neighbouring stops into account, it was able to conclude that the higher number of rides was a spatial pattern and it was not caused by the presence of a *secondary school*. The same principle goes for *car ownership*, *car ownership* is mostly found to be high in the villages and low in cities, in contradiction to the number of rides. The SEM concluded that the lower amount of rides was a spatial pattern and that it was not caused by the high car ownership.

The Moran I test for the SLX was significant, which means that the SLX model does not include all the independent variables that cause the spatial auto-correlation and it means that the chance of spatial auto-correlation in the error term is still high. To minimize the risk of false significance levels when explaining bus ridership, the SEM model should be interpret.

Much of the explanation power of the models can be attributed to frequency and therefore many of the 34 variables are redundant. Removing 22 variables resulted only in a reduction of explanatory power of 1.2% for the OLS model. The removed variables turned out to be slightly more important for the SLX model as the explanatory power was reduced with 1.5%. The remaining variables includes demographical, built environment and level of service variables and these variables will be used for the prediction model.

5 Predicting bus ridership

In chapter 3 three kinds of regression model were discussed. The OLS and SLX model turned out to be suitable for extrapolation. In this chapter the prediction accuracy of the two models, based on the twelve most influencing variables, is examined.

5.1 Prediction results

The spatial error distributions of the two models are examined, this way it possible to see if the prediction accuracy of the models within Arnhem-Nijmegen differs. The absolute values of the residuals are presented in figure 11 for the OLS and figure 12 for SLX. The colors are categorised in six groups in sizes of 2100 rides. This means, that for example the fourth group contains the the bus stops with residuals between 0 and 2100, in other words it includes the bus stops for which the model overestimates the actual number of rides, with a maximum of 2100. The limits have been set to multipliers of 2100, because there are 21 working days in November 2019, thus a residual of 2100 means that the model overestimates the actual number with 100 rides a day.

There is a clear distinction in the two figures, there are more bright colors in the figure 11, which means the OLS model has more large overestimates and large underestimates than the SLX model. Most of the brighter colors in figure 11 appear to be in the more rural areas in the cities Arnhem and Nijmegen contain more light colors.

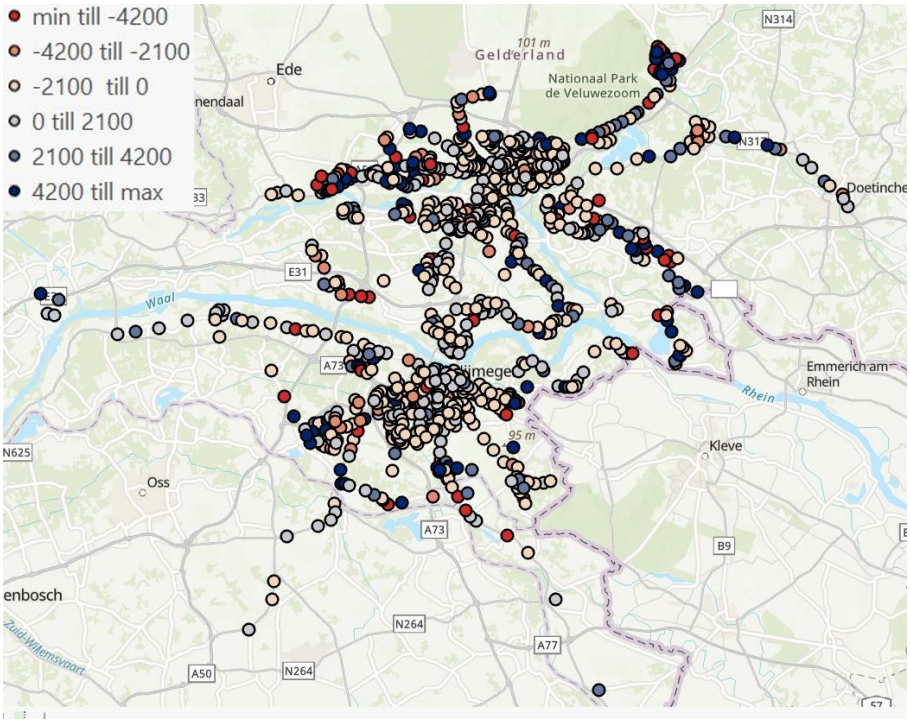

Figure 11: Residuals OLS

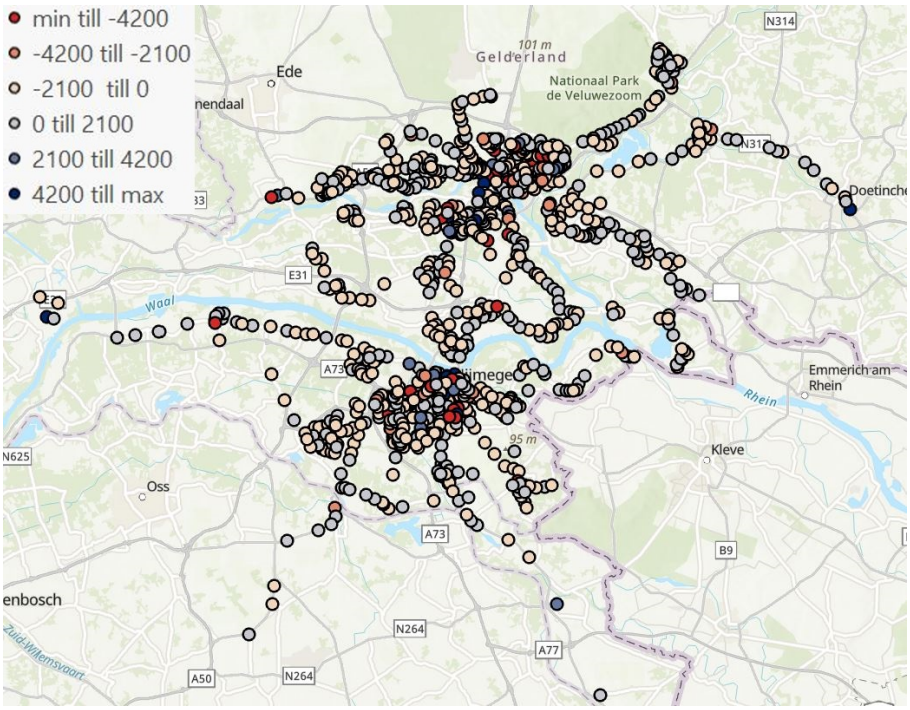

Figure 12: Residuals SLX

But, the absolute values do not tell the full story, as discussed in section 3.6. Therefore, the relative errors are displayed in figures 13 and 14 and they show a serious problem. For only a few stops, the models are able to predict between a -33% and 33% range of the actual value. Unfortunately, these stops are spread over the concession area, so it is not possible to determine an area for which the models perform well. For more than half of the bus stops, the predicted value of the SLX and OLS models was 50% more or 50% less than the actual value.

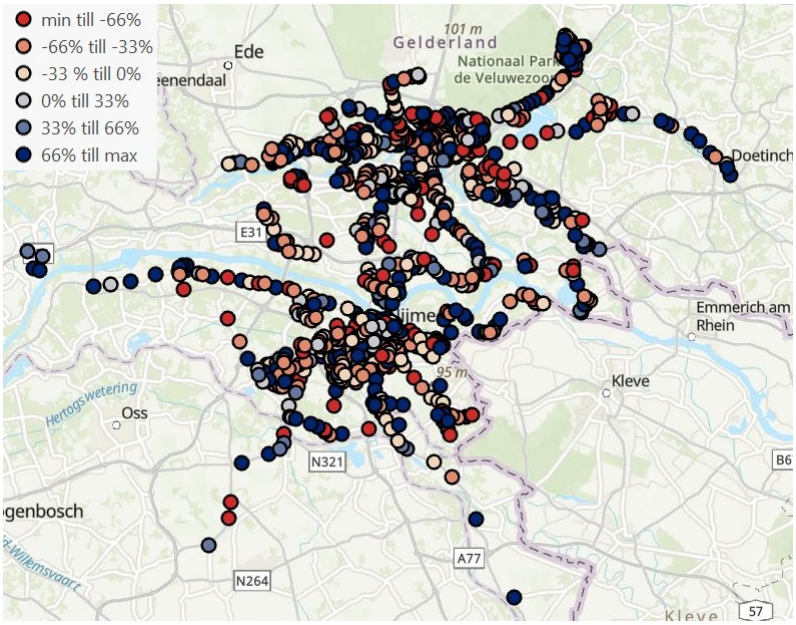

Figure 13: Relative residuals OLS

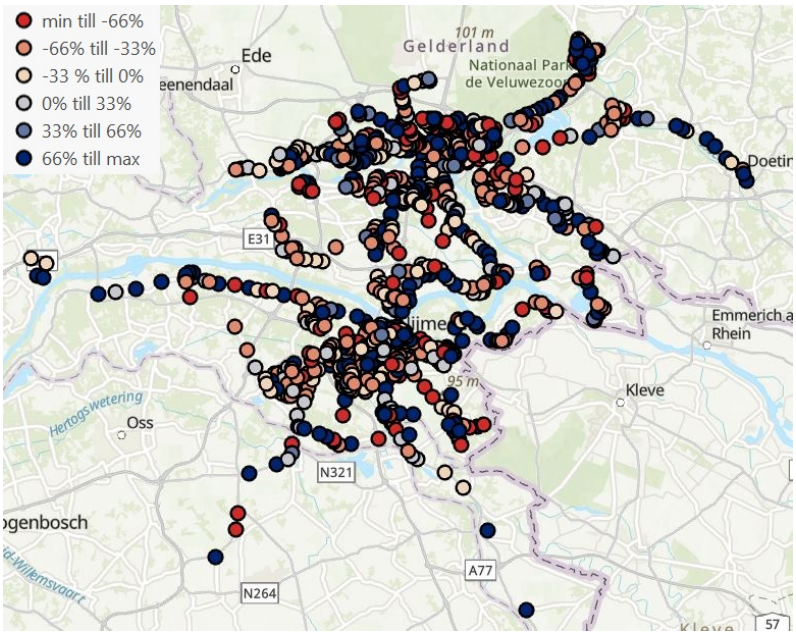

Figure 14: Relative residuals SLX

The difference between the predicted value and the actual can be best seen in figure 15. The axes are on a logarithmic scale, otherwise the majority of the stops would not be distinguishable because of the number of rides of Arnhem centraal. In the figure, it can be seen that the predictions for bus stops with 100 rides in November, range between 10 and 1000. For the stops with actual rides around the 10,000, it ranges from approximately, 1,000 to 12,000.

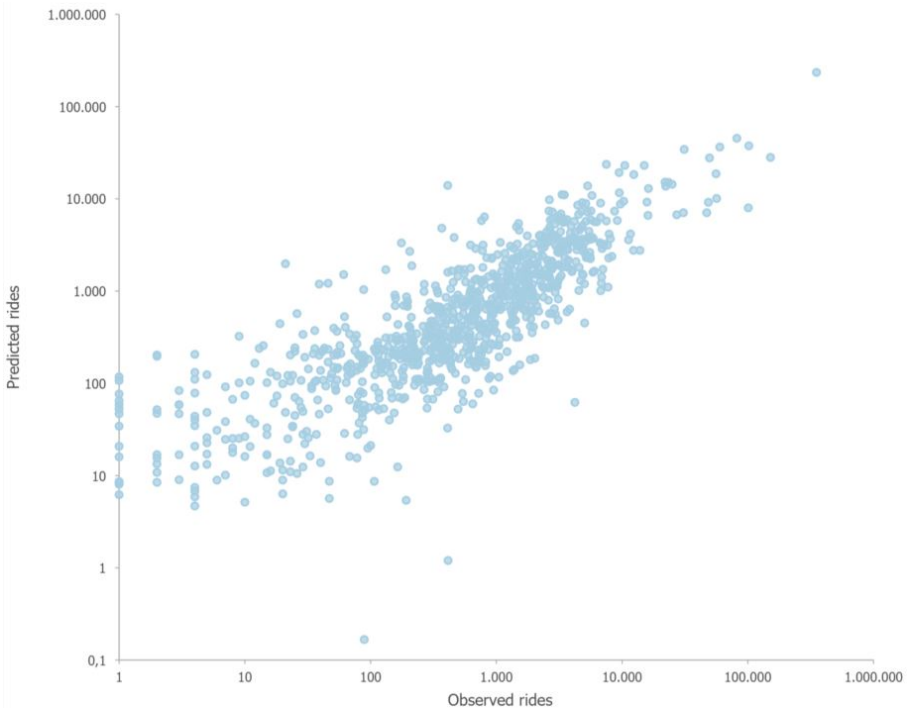

Figure 15: Scatter plot actual and predicted values

5.2 Prediction metrics

What was already slightly visible in the residual plots, is confirmed by looking at the prediction metrics, the SLX is the better performing model. For all the three metrics the SLX performs better, because it has more explanatory power and the Mean absolute error is smaller and the SLX have fewer extreme errors. In figure 16, the metrics are displayed again, but then categorised by the number of observed rides, to get a better idea of the sizes of the errors in comparison with the observed values. What can be seen is that the errors have substantive sizes. The OLS model seems to perform better for the group 100 to 1,000 rides and 1,000 and 10,000 rides.

Table 15: Prediction metrics, 16 independent variables

	R-squared	MAE	RMSE
OLS	70.2%	1594	8596
SLX	71.9%	1559	7527

Table 16: Prediction metrics categorised by observed rides

OLS	0 to 100	100 to 1,000	1,000 to 10,000	10,000 to 400,000
MAE	116	365	1,481	29,982
RMSE	255	817	2,212	48,883

SLX	0 to 100	100 to 1,000	1,000 to 10,000	10,000 to 400,000
MAE	107	381	1,557	27,749
RMSE	243	971	2,299	42,528

5.3 Conclusion

Unfortunately, this model is unable to make accurate predictions. The SLX model seem to perform better at extreme low values and extreme high values for rides, the OLS model seems to perform better for the rides from 100 to 10,000. But in both models the difference between the predicted value and the observed value is often more than 100% of the observed value.

In section 2.3.5, we saw in two studies that it is possible to make accurate predictions for ridership at stop level. What those two studies had in common was that the analysed stops could be considered as homogeneous. The studies contained only rapid transit stops which were located in urban areas.

In this study bus stops in urban and rural areas are examined and this resulted in high variances in values for the dependent and the independent variables. For instance, the standard deviation for rides in this study has a similar size as the standard deviation in the study of Guo and Huang (2020), namely respectively 14,199, see table 6, and 12,917. But the means for rides vary notably, 2,733 and 17,051.

The SLX model is able to deal with spatial auto-correlation in the independent variables, values for the independent variables are similar when they are close to each other. But the model is not capable to deal with non-stationary relationships between the ridership and the independent variables. It can be imagined that the problem of non-stationary is more severe when the study area contains bus stops in rural and urban area than when only stops in the same city are examined, as Guo and Huang (2020) and Cervero et al. (2010) did.

If non-stationary exists for bus ridership in the Netherlands, DRMs based on a study area cannot be generalised to other parts of the Netherlands. In the next chapter, it is examined if the relationships between the independent variables and ridership are the same for the area Arnhem-Nijmegen and for the area Groningen-Drenthe.

6 Analysing the possibility of generalizing the prediction model

The objective of the study is to develop a prediction model that is suitable for predictions of bus ridership at stop level for the Netherlands, with coefficients based on concession area Arnhem-Nijmegen. We have seen that the prediction accuracy of the model for its own study area is not satisfactory. To answer the fourth sub question and thus to see if it is possible at all to generalise the results of a DRM based on Arnhem-Nijmegen to other regions in the Netherlands, an OLS model will be estimated for the region Groningen-Drenthe. The results of the OLS model for Arnhem-Nijmegen and Groningen-Drenthe will be compared, to see if the areas are similar in terms of bus travel behaviour.

6.1 Study area Groningen-Drenthe

Smart card data is not easy to obtain in the Netherlands. OV Bureau Groningen-Drenthe provided smart card data for the region Groningen-Drenthe for the same month and year as the smart card data of Arnhem-Nijmegen. This makes it possible to compare bus travel behaviour between Arnhem-Nijmegen and Groningen-Drenthe and thus it makes the conclusion possible whether or not the prediction model based on Arnhem-Nijmegen can be applied to Groningen-Drenthe. The comparison is made by estimating an OLS model for Groningen-Drenthe, as the OLS model is the easiest method when the objective is explaining, with exactly the same independent variables and dependent variable. But first, we will have a look what kind of area Groningen-Drenthe is.

Where, the concession area Arnhem-Nijmegen contains only a part of the province of Gelderland, the concession area Groningen-Drenthe, as the name already suggests, completely contains the provinces Groningen and Drenthe. The size of the area is considerably larger than Arnhem-Nijmegen, 5,640 km2 against 1,100 km2. Groningen-Drenthe is a more rural area and has a more than a million residents (Alle cijfers, 2021). The population density is less than 200 residents per km2, less than a third of the density in Arnhem-Nijmegen. The biggest cities in Groningen-Drenthe are Groningen, Emmen and Assen, with respectively 200,000 (Allecijfers.nl, n.d.-c), 100,000 and 70,000 (Allecijfers.nl, n.d.-a) inhabitants.

Something similar for the two areas is the number of universities, both areas have one university. Furthermore, in Groningen-Drenthe the same public transport modes are available, the train and the bus. The area has 26 train stations and 1,671 bus stops. The differences between the two concession area makes Groningen-Drenthe a suitable area to validate if a DRM can be used to predict ridership in other parts of the Netherlands. What makes Groningen-Drenthe even more useful is the fact that Groningen-Drenthe is similar to a lot of different areas in the Netherlands. The provinces, Friesland, Flevoland, Overijsel and Zeeland have all similar population densities (metatopos, 2021).

Moreover, it is more useful to examine Arnhem-Nijmegen with a more rural area than with a more urban area. Because the Randstad (Amsterdam, Rotterdam, Den Haag and Utrecht) is the only area which is clearly more urban than Arnhem-Nijmegen. For the Randstad it can be assumed that the bus travel behaviour will be different. In Amsterdam and Rotterdam there are trams and metros. Den Haag has a tram and light rail and Utrecht has a tram. The availability of different public transit modes will probably lead to overestimating of the number of travellers for the bus.

The values for the descriptive statistics of rides and independent variables are displayed for Groningen-Drenthe. It is noticeable that the average number of rides per stop is a lot lower in Groningen-Drenthe than in Arnhem-Nijmegen, 1,295 to 2,732. Furthermore, the average amount of jobs than can be reached from a stop is about two thirds of the amount in Arnhem-Nijmegen.

Table 17: Descriptive statistics Groningen-Drenthe

	Unit	Minimum	Maximum	Mean	Std. Deviation
Rides	#/month	1.00	525320.00	1295.17	13507.67
Average car ownership	#/household	0.13	0.72	0.48	0.10
Elderly (65+)	%	1.80	66.15	21.38	8.14
LU residential	%	0.00	97.74	33.46	26.24
LU recreational	%	0.00	70.38	2.40	6.88
Distance to station	km	0.00	62.53	5.23	8.08
Hospitals within buffer	#	0.00	1.00	0.01	0.114
Higher education within buffer	#	0.00	1.00	0.01	0.09
Closest to train station	Dummy	0.00	1.00	0.02	0.12
Jobs accessible	#	0.00	118965.00	21230.20	22953.98
Hospitals accessible	#	0.00	3.00	0.59	0.666
Frequency	#/hour	0.04	105.82	3.11	4.61
Operators	#	1.00	4.00	1.05	0.25

6.2 OLS Groningen-Drenthe

The OLS model for Groningen-Drenthe is estimated with exactly the same variables as used as the reduced model for Arnhem-Nijmegen. After the model was estimated, the R-squared turned out to be a tentative indication that generalising the prediction model is going to be hard. The twelve independent variables can only explain only 57.5% of the ridership in Groningen-Drenthe, where the same variables could explain 70.2% in Arnhem-Nijmegen. And when the significance levels in table 18 are examined, it can be seen that only eight out of the twelve independent variables have a significance level lower than 0.1. The *percentage of elderly*, *distance to station*, *hospitals within buffer* and the *accessibility of hospitals* are not significant in Groningen-Drenthe.

To compare the relationships between the independent variables and the dependent variable, the standardized coefficients are examined. The signs of the standardized coefficients of the significant variables are the same for both areas. *Frequency* is affecting ridership the most for both areas, although there is a relatively large difference between the two coefficients. *Residential land-use*, for both the second most influential variable, turns out to be considerably more important in the provinces up north and the *number of operators* at a stop is more influential in Groningen-Drenthe than in Arnhem-Nijmegen. On the other side, *higher education within buffer*, the *accessibility of jobs*, *train stations* and *frequency* affect ridership more in Arnhem-Nijmegen.

Most of the difference between the standardized coefficients are likely to be caused by the vastness of Groningen-Drenthe. The car will be a more attractive transport mode in large areas of Groningen-Drenthe, because the travel times by car are a lot shorter than by bus. In the cities (with high *residential land-use*) the bus will be a more attractive option than in the rural areas (with low *residential land-use*). Because Arnhem-Nijmegen is more densely built, the bus will be more competitive with the car in rural areas than in Groningen-Drenthe, which explains the larger standardized coefficients for *residential land-use* for Groningen-Drenthe.

The same reasoning can be applied to *hospitals within buffer*. The *accessibility of hospitals* by bus is already lower in Groningen-Drenthe than in Arnhem-Nijmegen, see table 17 and because of the longer travel times, the bus is a less attractive option than the car when travelling to the hospital.

Table 18: Standardized coefficients OLS Arnhem-Nijmegen and Groningen-Drenthe

	Arnhem-Nijmegen		Groningen Drenthe	
	Std. Coefficient	Sig.	Std. Coefficient	Sig.
Average car ownership	-0.099	0.000***	-0.038	0.096*
Elderly (65+)	-0.074	0.000****	-0.002	0.914
LU residential	0.143	0.000***	0.318	0.000***
LU recreational	0.038	0.040**	0.028	0.087*
Distance to station	0.085	0.000***	-0.020	0.300
Hospitals within buffer	0.066	0.000***	0.019	0.244
Higher education within buffer	0.080	0.000***	0.042	0.011**
Closest to train station	0.066	0.000***	0.043	0.009***
Jobs accessible	0.071	0.003***	0.049	0.067*
Hospitals accessible	-0.062	0.004***	-0.005	0.794
Frequency	0.663	0.000***	0.557	0.000***
Operators	-0.036	0.064*	-0.079	0.000***

Significance levels: * 0.05 - 0.1, ** 0.01 - 0.05, *** <0.01

6.3 Conclusion

When the Netherlands is divided into three groups rural, urban and high urban. The rural and urban groups were assumed to have the most similar bus travel behaviour, because the bus and the train are the only forms of public transport. To see if a DRM based on an area in the Netherlands, could be used for predictions for a different area, the OLS results for Arnhem-Nijmegen and Groningen-Drenthe are compared. The twelve independent variables that could explain 70.2% of the bus ridership in Arnhem-Nijmegen, turned out to explain 57.5% in Groningen-Drenthe. Studying the standardized coefficients showed that the relationships between ridership and the independent variables differ significantly between the two areas. This is likely to be caused by the the car being the preferred transport mode in Groningen-Drenthe, since the difference in travel time between car and bus are large. In Arnhem-Nijmegen where in general the travel times are shorter, the bus is a more competitive transport mode for the car than in Groningen-Drenthe. Because of the differences in the relationships it is not advisable to generalize the model based on Arnhem-Nijmegen to different regions.

7 Applications of DRMs

A DRM is useful for public transport companies as it allows bus operators to efficiently allocate their resources. However in the studies mentioned in chapter 2 no detailed information is given on the applications of DRMs for policy makers, besides the interpretation of the results of the explaining purpose of DRMs. Because it is assumed DRMs have more applications for policy makers, in the next section three Dutch policy makers are interviewed for their view on applications of DRMs. Furthermore, in the introduction it was mentioned that the municipalities Arnhem and Nijmegen have put improvement of public transport high on the agenda. Therefore, after discussing the interviews, insights that can be drawn from the results in chapter 4 will be described and one application, which was identified over the course of the interviews, will be elaborated on for Arnhem-Nijmegen.

7.1 Interviews

The three policy makers are Mirjam Salomé, Erwin de Jager and Jorn van der Scheer. Salomé works at OV Oost, which is a public transport collaboration between the provinces Gelderland, Overijssel and Flevoland. In her role as coordinator of social safety role tries to ensure a feeling of safety for the public transport travellers. She is also involved in the process of planning new locations for bus stops located on provincial roads. Meanwhile Van der Scheer is a planner and transport developer for OV Bureau Groningen-Drenthe, a similar agency as OV Oost, and is involved in the design of bus routes. Finally, De Jager is a traffic engineer at the Groningen municipality and is involved in the execution phase of placing a new bus stop. As you can see the policy makers fulfil different roles in public transit sector. This should provide us with a number of different view points on the potential practical applications of DMs for policy makers.

One of the first questions during the interview with Salomé was: "What are the decisive criteria to determine new bus stop locations?". Salomé answered that the decision for a new bus stop is not a black-or-white issue, and that important factors are existing changing bus routes, road safety, the number of expected travellers and physical characteristics of the surroundings (such as protected nature areas or watercourses). When asked: "How is the number of expected travellers estimated?", Salomé answered that the surroundings, number of nearby houses and neighbouring facilities are all taken into account. To demonstrate the complexity of estimating the number of travellers, Salomé gave an example of an existing bus stop at a provincial road which was designed for six travellers a day. At one point a new neighbourhood was developed close to the stop and when the neighbourhood was finished, it led to much more travellers for the bus stop than it was designed for and this resulted in a dangerous traffic situation.

What Salomé described could have been prevented if a DRM was used. When during the construction phase of the neighbourhood, the new expected values for the independent variables of the bus stop were filled in the DRM, the DRM would show an increase in the number of rides in comparison with the actual number of rides. When this is done well before the neighbourhood was finished, there is enough time to adjust the bus stop. This shows that, besides DRMs allowing operators to adjust their services according to changes in neighbourhoods seen in section 1.1.3, they can also provide useful input for policy makers.

To find out if the ability to cope with changes has other applications, De Jager and Van der Scheer were asked the following question: DRMs are useful for coping with changes in the surroundings, but for what practical purposes is this a useful ability? De Jager and Van der Scheer came up with a different time related application of DRMs. They point out that bus stops and bus routes that exist for a long period of time are designed based on a couple assumptions, and a DRM could be useful to test whether these assumptions are still correct. For example they mentioned a recent trend in the Netherlands that is to no longer place bus stops in the center of neighbourhoods or villages but instead on the outskirts. This has resulted in shorter travel times as the bus does not have to go through a busy neighbourhood, and because of these shorter travel times the service frequency can be increased. When using a DRM it would be possible to examine if the number of travellers would increase when the bus stop is relocated to the outskirts.

In addition, DRMs according to Van der Scheer and de Jager can be used to identify the stops which have potential to increase the number of rides. If the predicted value of the DRM and the observed value differ significant, and assuming the DRM performs well, finding out why that stop does not live up to its potential, could lead to valuable insights on how the number of rides van be increased.

A DRM that includes accessibility variables could help policymakers with route planning. In Arnhem-Nijmegen, almost every bus route starts or ends at a (central) train station. In Groningen-Drenthe on the other hand they are experimenting with tangent lines, which connect the bus routes that are heading to the station. A bus system with tangent lines give travellers more opportunities to transfer or could eliminate the need for transfers altogether. When designing a tangent line, importing the new accessibility variables into the the DRM could give an insights in how many more travellers could be expected.

The general view of De Jager and Van der Scheer is on the DRM is that the DRM is useful to test hypotheses. They could for instance test if adding a new stop to a route leads to more travellers in total for a route. If the previous stop and the next stop of the route have notably less travellers than predicted then an extra stop may not be necessary. DRMs can also be used to quantify what for kind effect a bus stop has in a neighbourhood and this information can be used in the development phase of new neighbourhoods, by for example providing quick insights if the neighbourhood should have two or three bus stops. In addition, the explaining part of DRMs is considered always useful to get insights in travel behaviour.

Van der Scheer and De Jager emphasized that a DRM is useful for the reasons described above, but public transport decisions will not be based purely on the results of the DRMs. Just like Salomé, van der Scheer and de Jager point out that these kind of decisions are not always straightforward. Salomé mentioned the availability of space for a potential new bus stop location, while Van der Scheer and de Jager point out to the human aspect. Some households would be very happy if there is a plan to relocate a bus stop to the outskirts of a neighbourhood if that means that their children can play safely on the street.While others only see their access and egress time to the bus increasing and may thus try to block implementation/

7.2 Insights results explaining bus ridership Arnhem-Nijmegen

Only the relationships between the independent variables and bus ridership have been discussed in chapter 4. However the results in chapter 4 offer valuable insights for policy makers and those insights are discussed in this section. To interpret the results, the choice is made to follow the SEM model because of the proven spatial auto-correlation in the error terms in the OLS and SLX models. To start, the number of significant built environment and level of service variables, shows that policy makers can influence bus ridership. Also the positive relationship between parks and sport facilities has not been found in studies before. Policy makers could exploit this relationship by making it even more attractive to travel to recreational land use by bus, for instance by directing more bus routes to recreational areas. The same goes for amusement areas.

The positive relationship between hospitals within a catchment area and ridership was found in one study, but the relationship was not proven for Arnhem-Nijmegen before. It looks like here too the usage of buses can be increased when more bus routes are directed to hospitals.

Non-significant results are valuable as well. In chapter 2, two studies were discussed that concluded that having a P+R in the catchment area has a positive influence on the number of rides. But this relationship was found to be non-significant for P+R's in Arnhem-Nijmegen. P+R's can play an important role when it comes to reducing the number of cars in city centres, therefore the P+R's in Arnhem-Nijmegen should be made more attractive or should be promoted more.

7.3 Bus stops not fulfilling their potential

One of the applications of DRMs identified during the interviews is the ability to provide insights in which bus stop are not fulfilling their potential. To find the potential room for improvement in the number of rides of a bus stop, the observed value is subtracted from the value predicted by the SLX model. The SLX model is chosen over the OLS model because the SLX model performed better on the prediction metrics. In this section, five bus stops with high positive relative residuals in Arnhem-Nijmegen will be examined and some improvement suggestions will be provided.

7.3.1 Europalaan, Renkum

The bus stop Europalaan in Renkum has the highest relative residual in the subset of bus stops. 3,475 travellers have made use of the bus stop in November 2019 while the SLX model predicted this value would be 11,050. In figure 16, the catchment area of bus stop Europalaan is shown, as well as the location of other bus stops in the region. The residential land use in the surroundings of Europalaan is very high (90%), which explains the high number of predicted rides.

The catchment area has a 400 meter radius, but this distance is as the crow flies. Because of the design of the neighbourhood and the location of the stop, the walking distance between locations in the catchment area to the stop is often more than 400 meter. To make the bus stop better accessible for residents and thus to make the bus stop more attractive, the bus stop could be relocated. On the south end of Europalaan there are many competing stops, so the advice is to examine if it is possible to relocate the Europalaan stop to the north, preferably close to an intersection. This way the walking distance to the stop for many residents is reduced. The increase in travel time for the bus will be minimal, as the current location of the stop and the new location will be both in the middle of a residential neighbourhood.

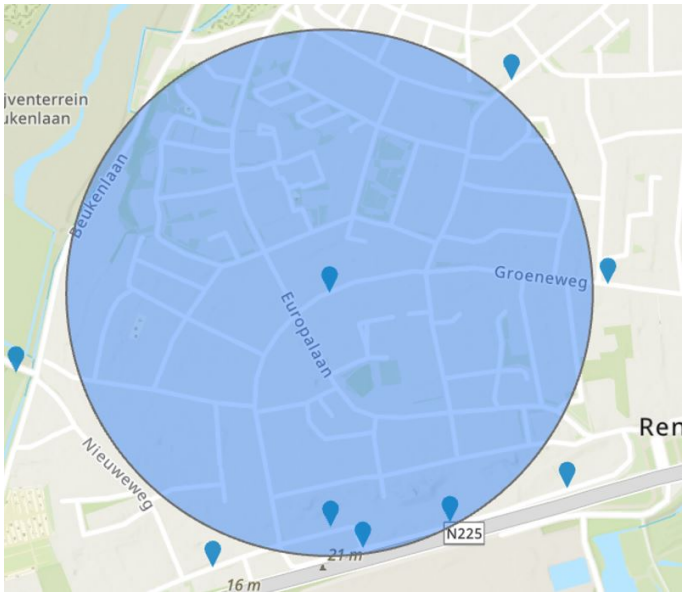

Figure 16: Busstop Europalaan, Renkum

7.3.2 Beemdstraat, Nijmegen

The stop Beemdstraat is located in a residential neighbourhood in Nijmegen. The number of observed rides is 2654 while the number of predicted rides is 4916. In comparison with the other bus stops in the catchment area (6,741, 5,446 and 3,246 rides), the number of observed rides rides for Beemdstraat is low. This is likely caused by the low number of directions (two) of the Beemdstraat. The directions of two of the other stops are five and seven and as we have seen in

the results of the SLX-model (section 4.4.2), more directions can attract travellers from stops with lower directions.

It seems that the stop Beemdstraat may be be redundant. When the stop is removed other bus stops are still reachable within 400 meters walking distance for all residents. As a bonus the travel time of the bus will be reduced making it more attractive to take the bus. The increase in walking distance for some of the residents will is mostly a problem for the elderly living in their neighbourhood. But fortunately this will not be severe problem as the percentage of elderly (10%) is relatively low.

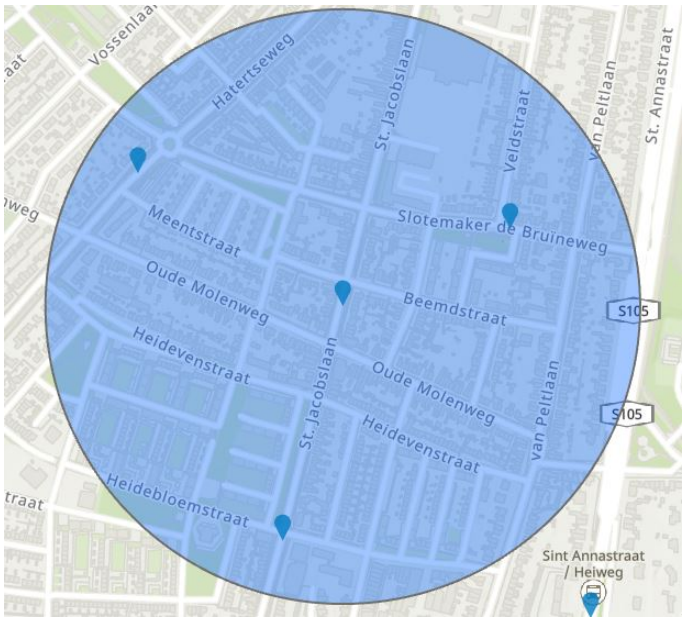

Figure 17: Bus stop Beemdstraat, Nijmegen

7.3.3 Kerkstraat, Weurt

Weurt is a small town that lies to the west of Nijmegen that has one bus stop, named Kerkstraat. Kerkstraat has 2,810 rides per month, while 6,086 is the predicted number of rides. For this bus stop it is hard to determine a way to increase the number of rides. The stop lies almost in the centre of Weurt and an inspection using Google Maps Street View shows that traffic lights make that the bus stop is easily accessible. The city centre of Nijmegen can be reached by bus within eight minutes, which is only two minutes slower than by car. The average car ownership in the catchment area is with 1.14 rather high. In this case the advice would therefore be to actively promote bus use in the surrounding neighbourhood, highlighting the only minimal difference in travel time and maybe the positive environmental impact as well. Time will tell if this convinces the inhabitants of Weurt to take the bus instead of the car.

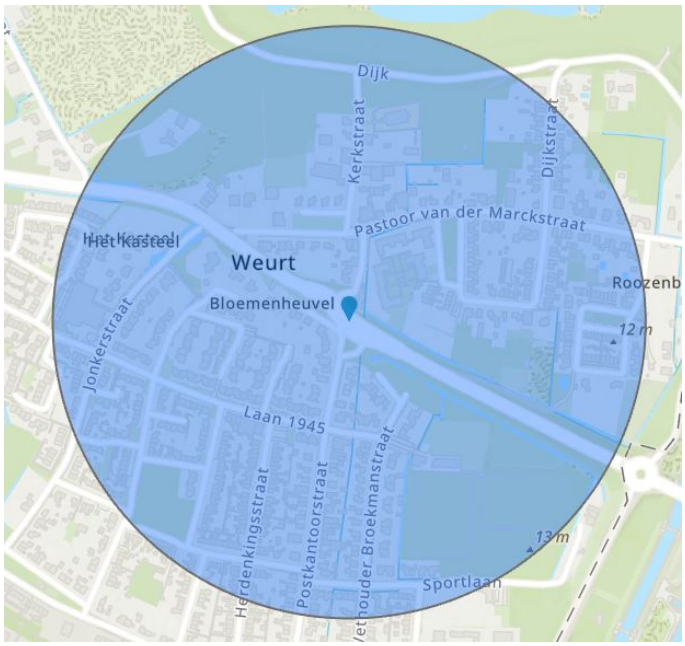

Figure 18: Bus stop Kerkstraat, Weurt

7.3.4 Fransestraat, Nijmegen

For the bus stop Fransestraat in Nijmegen the model predicted a number of rides of 4,916 while the actual value is 2,654, resulting in a relative residual of 85%. In comparison with the means for the independent variables seen in table 6 the residential land use (72%) is high and the percentage of elderly (8%) is very low, which are both causes of the high predicted value. Interesting to see is that bus stops on the same street, with the same bus lines are performing better in terms of the relative residual. The two closest bus stops have relative residuals around 30%. Differences in the surrounding of the stops and Fransestraat are the high percentage of youth and the very low car ownership in the surroundings of Fransestraat.

The high percentage of youth and low car ownership are clear indicators of many students living close to the Fransestraat stop, but the bus routes are not passing higher education facilities. It is likely that redirecting a bus route to pass the Fransestraat and a higher education facility will increase the number of rides, but this will probably be at the expense of bike rides in stead of car rides. Since this is not the goal of the municipality the advice for this stop is to leave the situation as is.

7.3.5 Hatertseweg, Nijmegen

For the stop Hatertseweg in Nijmegen more than 6,000 rides were predicted while only 3,393 were observed. From this stop it takes 24 minutes by bus to reach the city centre of Nijmegen, which is 15 minutes longer than travelling by car. The large difference between is caused by the 10 minute walking time from Nijmegen central station to the city centre, as the train station is the final stop for the bus lines passing Hatertseweg. If the bus routes are extended and they are designed to drive through and stop in the city centre as well, taking the bus to the center could become a more attractive alternative.

7.4 Conclusion

Interviews with Salomé, Van der Scheer and de Jager showed that a model can predict bus ridership will not be the only decisive factor in decisions on the location of bus stops, because there are many issues that should be considered such as the local opinion on a bus stop and the availability of space. But a DRM is considered as a useful tool to estimate the effects of changes in the surroundings of a bus stop. For instance, adjustments to an existing bus stop can be made in time when nearby a neighbourhood is developed. Another application is to identify stops which are performing lower than expected. The general view on DRMs is that they can help substantiating decisions about planning new stops and relocating existing stops and they can provide useful information for designing new bus routes.

The explaining part of the DRM showed that policy makers could influence bus ridership as bus ridership is significantly related to built environment and level of service variables. To encourage bus ridership in Arnhem-Nijmegen, it could be examined if transportation by bus to hospitals, parks, amusement parks and sport facilities could be facilitated more. Furthermore, it could be examined why P+R's do not have a significant relationship with ridership.

The prediction part of a DRM was used to identify five bus stops in Arnhem Nijmegen, that were performing lower than expected. Although it could in fact be the case that the predictions are inaccurate and that the bus stops are not under performing, suggestions how to cope with under performing stops are provided, which could be applied to other bus stops as well.

8 Conclusion

The objective of this study was to develop a model, based on the concession area Arnhem-Nijmegen, that could give a first insight in the number of travellers at a bus stop in the Netherlands. This objective was translated into the following research question: "What is the influence of variables on bus ridership at stop level and what is the predictive accuracy and the usability of the model for policymakers?"

To answer this question, the first step was conducting a literature review of variables which are known to affect bus ridership and of the methods used to explain and predict bus ridership. Demographics, built environment and level of service variables turned out to be the most used variables and a Direct Ridership Model was considered to be the best method to explain and predict bus ridership. Advantages of the DRM are the ability to evaluate new explanatory variables, high transparency, high explanation power and the fact that it allows to experiment with changes in the explanatory variables.

Three DRMs were estimated based on three different regression models, namely Ordinary Least Squares, a Spatial Lag X and Spatial Error model. The latter two can take spatial auto-correlation into account and proved to be a valuable addition compared to the OLS model. The SLX turned out to be the most useful for predicting bus ridership as it is possible to make predictions out of the study area. The SEM model was considered to be the best model for explaining bus ridership, because the spatial auto-correlation in the residuals of the OLS and SLX models could lead to false significant values. The SEM showed, in contrast to the OLS model, that the relationship between bus ridership and car ownership is not significant. Other new insights on explaining bus ridership in the Netherlands are that average household size, the percentage of 15-24 year olds, the distance to station, recreational land-use (parks and sport facilities) and amusement land-use (amusement parks) were found to positively affect bus ridership. No significant relationships were found for political preference, the presence of secondary schools, the presence of bicycle stand at a bus stop and the accessibility of jobs, hospitals and higher education from the bus stop.

To examine the prediction accuracy the estimates of the OLS and SLX model were compared with the observed values in Arnhem-Nijmegen. The accuracy of the SLX model turned out to be better than the accuracy of the OLS model, but in general the prediction accuracy of both models turned out to be poor. The low prediction accuracy is likely caused by varying relationships between ridership and the independent variables in urban and rural areas. The term for spatial varying relationships is non-stationary.

That non-stationary relationships exists for bus ridership was proved, while the study looked into if a ridership prediction model based on Arnhem-Nijmegen could be generalised to different regions in the Netherlands. Two OLS models were estimated, one was based on Arnhem-Nijmegen and the other on the concession area Groningen-Drenthe. It turned out there were significant differences in the sizes of the standardized coefficients between the regions. Therefore it is not advisable to generalize the prediction model based on Arnhem-Nijmegen.

Furthermore demonstrating the added value of spatial relationships for explaining and predicting bus ridership and demonstrating DRMs cannot be generalised to different regions because of non-stationary relationships, this study identified applications of DRMs for policy makers. For example, when a construction on a new residential neighbourhood is finished, it could lead to more travellers using a bus stop than the stop was designed for and this could result in dangerous traffic situations. With a DRM it is possible to predict the number of travellers before the neighbourhood is finished and this allows the bus stop to be adjusted accordingly. Another application is identifying bus stops for which the actual number of travellers is far lower than the predicted number. The general application of a DRM is providing substantiation for plans for new routes, new stop locations and relocating bus stops.

For the region Arnhem-Nijmegen five bus stops have been identified which are under performing, according to the model. To increase the number of rides for these bus stops suggestions, such as relocating the bus stop or redirecting the bus route, are provided.

This study is based on several assumptions and most of them are related to data. To calculate the values of demographics of a catchment areas it is assumed that for instance the number of residents in a zip code area is evenly distributed. If 25% of a zip code area of 1,000 residents lies in a catchment area of a bus, it is assumed 250 resident live in the overlap of the zip code area and the catchment area. Furthermore, because the objective was to develop a model that can be applied nation-wide it is assumed that the closest bus stop to a train station is the main bus stop of that station. This way, linking train station to the closest bus stop could be automated. Other data related assumptions are: the assumption that travellers are willing to take bus rides with with a maximum duration of 45 minutes or that the influence of overlapping catchment areas can be ignored when the stops are more than 400 meters apart from each other.

One assumption related to methodology is that the ridership of every bus stop is influenced by the three closest stops. When a bus stop is the only stop in a village, the influence of neighbouring stops can be questioned. Furthermore, is is assumed that the twelve most explaining variables are the same for the OLS and SLX models.

For the explanatory results, it is assumed the spatial auto-correlation in the residuals in the SEM model is zero. But the presence of spatial auto-correlation is not ruled out. A further concern related to the results is the endogeneity problem, which is likely to have caused an overestimation of the independent variables frequency and number of directions.

This study can serve as a basis for further research. There are various ways how the results of this study can be improved. For further research it is recommended to have a smaller more homogeneous study area, for example only bus stops in cities. It would be interesting to see if a DRM based on a city can be generalised to different cities. When the only objective is explaining the complete area Arnhem-Nijmegen can be analysed with a geographically weighted regression model, as this model is able to deal with non-stationary relationships.

Other improvements could be distance decayed catchment areas based on walking distance and a spatial weight matrix based on distance. Furthermore, extra explanatory variables can be included such as number of residents instead of residential land-use or variables that encourage car use instead of bus use such as the difference in travel time to the city center between car and bus or the distance to the closest highway ramp.

References

Aggarwal, V., Gupta, V., Singh, P., Sharma, K., & Sharma, N. (2019). Detection of spatial outlier by using improved z-score test. In *2019 3rd international conference on trends in electronics and informatics (icoei)* (pp. 788–790).

Alle Cijer. (2021). *Informatio gemeente venlo.* Retrieved from https://allecijfers.nl/gemeente/venlo/

Alle cijfers. (2021). *Nederland.* Retrieved from https://allecijfers.nl/provincie/groningen/

Allecijfers.nl. (n.d.-a). *De grootste gemeenten in inwoners in drenthe:.* Retrieved from https://allecijfers.nl/ranglijst/de-grootste-gemeenten-in-inwoners-in-drenthe/

Allecijfers.nl. (n.d.-b). *De grootste gemeenten in inwoners in gelderland.* Retrieved from https://allecijfers.nl/ranglijst/de-grootste-gemeenten-in-inwoners-in-gelderland/

Allecijfers.nl. (n.d.-c). *Informatie gemeente groningen:.* Retrieved from https://allecijfers.nl/gemeente/groningen/

Allison, P. (2014). Prediction vs. causation in regression analysis. Retrieved from https://scholar.googleusercontent.com/scholar.bib?q=info:jJbRN8WxIrIJ:scholar.google.com/&output=citation&scisdr=CgVUAS9KEM7KzydHCXI:AAGBfm0AAAAAYKZBEXJL1vjpxuIIwGcjj-ugC2u-rL0H&scisig=AAGBfm0AAAAAYKZBEZhii9lrt6r6sM4Y9iMAGul73YGm&scisf=4&ct=citation&cd=-1&hl=en

Anselin, L. (2002). Under the hood issues in the specification and interpretation of spatial regression models. *Agricultural economics*, *27*(3), 247–267.

Anselin, L. (2003). Spatial externalities, spatial multipliers, and spatial econometrics. *International regional science review*, *26*(2), 153–166.

Anselin, L., & Bera, A. K. (1998). Introduction to spatial econometrics. *Handbook of applied economic statistics*, *237*.

Arnhem2day. (n.d.). *De trolleybus.* Retrieved from https://www.arnhem2day.nl/fotobank/trolleybus ((Accessed on 05/04/2021))

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational research methods*, *7*(2), 127–150.

Beirão, G., & Cabral, J. S. (2007). Understanding attitudes towards public transport and private car: A qualitative study. *Transport policy*, *14*(6), 478–489.

Bouma, F., & Bontjes, A. (2019). *Meer omgebouwde huurwoningen voor jonge alleenstaanden.* Retrieved from https://www.nrc.nl/nieuws/2019/10/22/meer-omgebouwde-huurwoningen-voor-jonge-alleenstaanden-a3977594

Boussiala, M. (2020, 10). *Cook's distance.* doi: 10.13140/RG.2.2.18888.55049

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243.

Bremmer, D. (2020). *Krijgen nieuwbouwwoningen te weinig parkeerplekken?* Retrieved from https://www.ad.nl/binnenland/nieuwbouwhuis-dan-kun-je-een-parkeerplek-wel-vergeten~a8a461e3/#:~:text=De%20normen%20maken%20onderscheid%20in,norm%20van%200%2C2%20geldt.

Burkey, M. L. (2018). Spatial econometrics and gis youtube playlist. *REGION*, *5*(3), R13–R18.

Cardozo, O. D., García-Palomares, J. C., & Gutiérrez, J. (2012). Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography*, *34*, 548–558.

CBS. (2017). *Kerncijfers per postcode.* https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode. ((Accessed on 04/28/2021))

CBS. (2018). *Kerncijfers per postcode.* https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode. ((Accessed on 04/28/2021))

CBS. (2019). *Huishoudens naar inkomen en autobezit per pc5-gebied.* https://www.cbs.nl/nl-nl/maatwerk/2019/23/huishoudens-naar-inkomen-en-autobezit-per-pc5-gebied.

Cervero, R., Murakami, J., & Miller, M. (2010). Direct ridership model of bus rapid transit in los angeles county, california. *Transportation Research Record*, *2145*(1), 1–7.

Chai, T., & Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, *7*(3), 1247–1250.

Chiou, Y.-C., Jou, R.-C., & Yang, C.-H. (2015). Factors affecting public transportation usage rate: Geographically weighted regression. *Transportation Research Part A: Policy and Practice*, *78*, 161–177.

Chow, L.-F., Zhao, F., Liu, X., Li, M.-T., & Ubaka, I. (2006). Transit ridership model based on geographically weighted regression. *Transportation Research Record*, *1972*(1), 105–114.

Chu, X. (2004). Ridership models at the stop level. *Report No. BC137-31 prepared by National Center for Transit Research for Florida Department of Transportation*.

Chu, X., Polzin, S. E., Pendyala, R. M., Siddiqui, N. A., & Ubaka, M. (2006). *Framework of modeling and forecasting stop-level transit patronage* (Tech. Rep.).

Clifton, K. J., & Handy, S. L. (2001). *Qualitative methods in travel behaviour research.* Citeseer.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18.

Costa, V., Fontes, T., Costa, P. M., & Dias, T. G. (2015). Prediction of journey destination in urban public transport. *Progress in Artificial Intelligence Lecture Notes in Computer Science*, 169–180. doi: 10.1007/978-3-319-23485-4_18

CPB and KIM. (2009). *Het belang van openbaar vervoer.* Retrieved from `https://www.cpb.nl/sites/default/files/publicaties/download/het-belang-van-openbaar-vervoer-de-maatschappelijke-effecten-op-een-rij.pdf`

Crow. (2018). *Krijgen nieuwbouwwoningen te weinig parkeerplekken?* Retrieved from `https://www.crow.nl/over-crow/nieuws/2018/juli/krijgen-nieuwbouwwoningen-te-weinig-parkeerplekken`

Crow. (2020). *Regionaal openbaar vervoer per 1 januari 2020.* Retrieved from `https://www.crow.nl/downloads/documents/2019/concessieposter2020.aspx?ext=.pdf`

Dajani, J., & Sullivan, D. (1976). A causal model for estimating public transit ridership using census data. *High Speed Ground Transportation Journal*, *10*(1).

Davarzani, N., Peeters, R., Smirnov, E., Karel, J., & Brunner-La Rocca, H.-P. (2016). Ranking accuracy for logistic-gee models. In *International symposium on intelligent data analysis* (pp. 14–25).

Den Boer, L. C., & Vermeulen, J. P. L. (2004, December). *Snelheid en milieu.* Retrieved from `https://www.ce.nl/publicaties/download/245`

Dill, J., Schlossberg, M., Ma, L., & Meyer, C. (2013). Predicting transit ridership at the stop level: The role of service and urban form. In *92nd annual meeting of the transportation research board, washington, dc.*

DUO. (2021). *Onderwijslocaties adres.* Retrieved from `https://services.arcgis.com/nSZVuSZjHpEZZbRo/ArcGIS/rest/services/DUO_Onderwijslocaties_1/FeatureServer/0`

Durning, M., & Townsend, C. (2015). Direct ridership model of rail rapid transit systems in canada. *Transportation Research Record*, *2537*(1), 96–102.

Eurostat. (n.d.). *Population demography migration projections.* Retrieved from `https://ec.europa.eu/eurostat/en/web/population-demography-migration-projections/statistics-illustrated`

Folkert, N. (2019). *Internationaal gezien valt de verkeersdrukte in nederland mee.* `https://www.nt.nl/wegvervoer/2019/06/12/internationaal-gezien-valt-de-verkeersdrukte-in-nederland-mee/?gdpr=accept`. ((Accessed on 04/29/2021))

Gan, Z., Feng, T., Yang, M., Timmermans, H., & Luo, J. (2019). Analysis of metro station ridership considering spatial heterogeneity. *Chinese Geographical Science*, *29*(6), 1065–1077.

Gemeente Amsterdam. (2013). *Amsterdam aantrekkelijk bereikbaar.* Retrieved from `https://assets.amsterdam.nl/publish/pages/865232/mobiliteitsaanpak_amsterdam_2030.pdf`

Gemeente Arnhem. (2016). *Beleidsnota openbaar vervoer.* Retrieved from https://www.arnhem6811.nl/wp-content/uploads/2016/08/Beleidsnota-Openbaar-Vervoer.pdf

Gemeente Nijmegen. (2018). *Ambitiedocument mobiliteit 2019-2030.* Retrieved from https://www.slimschoononderweg.nl/wp-content/uploads/2019/07/AmbitiedocumentMobiliteitNijmegen.pdf

Gemeente Utrecht. (2020). *Mobiliteitsplan 2040.* Retrieved from https://omgevingsvisie.utrecht.nl/fileadmin/uploads/documenten/zz-omgevingsvisie/thematisch-beleid/verkeer-mobiliteit/2021-05-mobiliteitsplan-2040-toegankelijk.pdf

Gimenez-Nadal, J. I., & Molina, J. A. (2014). Commuting time and labour supply in the netherlands a time use study. *Journal of Transport Economics and Policy (JTEP)*, *48*(3), 409–426.

Goulard, M., Laurent, T., & Thomas-Agnan, C. (2017). About predictions in spatial autoregressive models: Optimal and almost optimal strategies. *Spatial Economic Analysis*, *12*(2-3), 304–325.

Griffith, D. A. (2000). A linear regression solution to the spatial autocorrelation problem. *Journal of Geographical Systems*, *2*(2), 141–156.

Grosvenor, T. (2000). *Qualitative research in the transport sector* (Tech. Rep.).

Guo, R., & Huang, Z. (2020). Mass rapid transit ridership forecast based on direct ridership models: A case study in wuhan, china. *Journal of Advanced Transportation*, *2020*.

Gutiérrez, J., Cardozo, O. D., & García-Palomares, J. C. (2011). Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. *Journal of Transport Geography*, *19*(6), 1081–1092.

Hensher, D. A. (2008). Assessing systematic sources of variation in public transport elasticities: some comparative warnings. *Transportation Research Part A: Policy and Practice*, *42*(7), 1031–1042.

IBM. (n.d.). *Linear regression variable selection methods.* Retrieved from https://www.ibm.com/docs/en/spss-statistics/23.0.0?topic=regression-linear-variable-selection-methods

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation review*, *5*(5), 602–619.

Kerkman, K., Martens, K., & Meurs, H. (2015). Factors influencing stop-level transit ridership in arnhem–nijmegen city region, netherlands. *Transportation Research Record*, *2537*(1), 23-32. Retrieved from https://doi.org/10.3141/2537-03 doi: 10.3141/2537-03

Kerkman, K., Martens, K., & Meurs, H. (2018). Predicting travel flows with spatially explicit aggregate models: on the benefits of including spatial dependence in travel demand modeling. *Transportation Research Part A: Policy and Practice*, *118*, 68–88.

Kiesraad. (n.d.). *Verkiezingsuitslag tweede kamer 2021 — data overheid.* https://data.overheid.nl/dataset/verkiezingsuitslag-tweede-kamer-2021. ((Accessed on 04/30/2021))

Kikuchi, S., & Miljkovic, D. (2001). Use of fuzzy inference for modeling prediction of transit ridership at individual stops. *Transportation research record*, *1774*(1), 25–35.

KIM. (2018). *Cycling facts.* Retrieved from https://www.government.nl/binaries/government/documents/reports/2018/04/01/cycling-facts-2018/Cycling+facts+2018.pdf

KiM. (2018). *De keuze van de reiziger.* Retrieved from https://www.kimnet.nl/binaries/kimnet/documenten/rapporten/2018/05/07/documentatie-de-keuze-van-de-reiziger/Keuze+van+de+reiziger.pdf

KIM. (2019). *Mobiliteitsbeeld 2019.* Retrieved from https://www.kimnet.nl/publicaties/rapporten/2019/11/12/mobiliteitsbeeld-2019-vooral-het-gebruik-van-de-trein-neemt-toe

Kloosterboer, D. (2021). *Kiesraad.* Retrieved from https://github.com/DIRKMJK/kiesraad/tree/master/kiesraad

Koops, R. (2019). *Ruim 10.000 parkeerplaatsen verdwijnen voor 2025.* Retrieved from https://www.parool.nl/nieuws/ruim-10-000-parkeerplaatsen-verdwijnen-voor-2025~b8496335/

Kuby, M., Barranda, A., & Upchurch, C. (2004). Factors influencing light-rail station boardings in the united states. *Transportation Research Part A: Policy and Practice*, *38*(3), 223–247.

Kuiken, A. (2020, December 28). *Verkeer in steden dreigt vast te lopen.* Retrieved from https://www.trouw.nl/nieuws/verkeer-in-steden-dreigt-vast-te-lopen~bec66e87/#:~:text=Vertaald%20naar%20de%20steden%20leidt,stijgt%20het%20aantal%20files%20exponentieel

Lindner, I. (2013). *Statistics ii.* Retrieved from https://personal.vu.nl/i.d.lindner/Tutorial%201.PDF

Liu, Z. (1993). Determinants of public transit ridership analysis of post world war ii trends and evaluation of alternative networks.

Luo, H., Zhang, Z., Gkritza, K., & Cai, H. (2021). Are shared electric scooters competing with buses? a case study in indianapolis. *Transportation Research Part D: Transport and Environment*, *97*, 102877.

Lynch, S. M., & Brown, J. S. (2011). Stratification and inequality over the life course. In *Handbook of aging and the social sciences* (pp. 105–117). Elsevier.

MacKenzie, S. B. (2001). Opportunities for improving consumer research through latent variable structural equation modeling. *Journal of Consumer Research*, *28*(1), 159–166.

Marquaridt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, *12*(3), 591–612.

McNally, M. G. (2007). *The four-step model*. Emerald Group Publishing Limited.

metatopos. (2021). *Provincies nederland.* Retrieved from https://www.metatopos.eu/provincies_eu.php

Morris, J. D., & Lieberman, M. G. (2018). Multicollinearity's effect on regression prediction accuracy with real data structures. *General Linear Model Journal*, *44*, 29–34.

MSU. (n.d.). *Hypothesis testing, effect size, power.* Retrieved from http://web.mnstate.edu/malonech/Psy633/Notes/hyp.%20testing,%20effect%20size,%20power%20GW8.htm#:~:text=Also%2C%20if%20the%20sample%20size,to%20reject%20the%20null%20hypothesis.&text=The%20appearance%20of%20a%2015,point%20effect%20is%20relatively%20small. ((Accessed on 05/04/2021))

Mucci, R. A., & Erhardt, G. D. (2018). Evaluating the ability of transit direct ridership models to forecast medium-term ridership changes: Evidence from san francisco. *Transportation Research Record*, *2672*(46), 21–30.

Murack, J. (2013). *Regression analysis using gis.* https://libraries.mit.edu/files/gis/regression_presentation_iap2013.pdf. ((Accessed on 04/29/2021))

NDOV. (n.d.). *Nationaal halte bestand.* ://data.ndovloket.nl/haltes/.

Niaounakis, T. K., Blank, J. L., & Veeneman, W. (2016). Doelmatig aanbesteden.

Nunkoo, R., & Gursoy, D. (2012). Residents' support for tourism: An identity perspective. *Annals of tourism research*, *39*(1), 243–268.

Nunkoo, R., & Ramkissoon, H. (2012). Structural equation modelling and regression analysis in tourism research. *Current Issues in Tourism*, *15*(8), 777–802.

Omroep Gelderland. (2019). *Arnhem en nijmegen hebben ov-ambities maar wachten op gelderse miljoenen.* Retrieved from https://www.omroepgelderland.nl/nieuws/2409610/Arnhem-en-Nijmegen-hebben-OV-ambities-maar-wachten-op-Gelderse-miljoenen

Open mobility data. (2019). *14 november 2019.* Retrieved from https://transitfeeds.com/p/ov/814/20191114

Open State Foundation. (n.d.). *Stembureaus tweede kamerverkiezingen 2021.* ://data.overheid.nl/dataset/stembureaus-tweede-kamerverkiezingen-2021.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, *9*(1), 6.

Peel, M. J., Goode, M. M., & Moutinho, L. A. (1998). Estimating consumer satisfaction: Ols versus ordered probability models. *International Journal of Commerce and Management*.

Prairie, Y. T. (1996). Evaluating the predictive power of regression models. *Canadian Journal of Fisheries and Aquatic Sciences*, *53*(3), 490–492.

Propastin, P., Kappas, M., & Erasmi, S. (2008). Application of geographically weighted regression to investigate the impact of scale on prediction uncertainty by modelling relationship between vegetation and climate. *International journal of spatial data infrastructures research*, *3*(3), 73–94.

Puentes, R. (2016). *Why infrastructure matters: Rotten roads, bum economy.* Retrieved from https://www.brookings.edu/opinions/why-infrastructure-matters-rotten-roads-bum-economy/#:%7E:text=The%20economy%20needs%20reliable%20infrastructure,transit%20can%20reduce%20greenhouse%20gases

Pulugurtha, S. S., & Agurla, M. (2012). Assessment of models to estimate bus-stop level transit ridership using spatial modeling methods. *Journal of Public Transportation*, *15*(1), 3.

RDW. (2021a). *Open data parkeren.* Retrieved from https://opendata.rdw.nl/browse?category=Parkeren&limitTo=datasets

RDW. (2021b). *Open data parkeren.* Retrieved from https://opendata.rdw.nl/browse?category=Parkeren

Regenmaker. (2019). *Concessies nederland nummers.* Retrieved from https://wiki.ovinnederland.nl/wiki/Bestand:Concessies_Nederland_nummers.png#filelinks

RHDHV. (2019). *Bodemgebruik 2015 wfl1.* Retrieved from https://www.arcgis.com/home/item.html?id=431475e0006542609fc2e577b9c17932

Rijksoverheid. (n.d.). *Afspraken over regionaal en stedelijk openbaar vervoer.* Retrieved from https://www.rijksoverheid.nl/onderwerpen/openbaar-vervoer/afspraken-over-het-openbaar-vervoer/afspraken-over-regionaal-openbaar-vervoer

Ryan, S., & Frank, L. (2009). Pedestrian environments and transit ridership. *Journal of Public Transportation*, *12*. doi: 10.5038/2375-0901.12.1.3

Sale, J. (1976). *Increasing transit ridership: the experience of seven cities.* Urban Mass Transportation Administration.

Shepherd, B. (2009). *Session 3: Dealing with endogeneity.* Retrieved from https://artnet.unescap.org/tid/artnet/mtg/gravity09_tues3.pdf

Shmueli, G. (2010). To explain or to predict? *Statistical science*, *25*(3), 289–310.

Sincich, T. (1993). *Statistics by example.* Dellen.

Sohn, K., & Shim, H. (2010). Factors generating boardings at metro stations in the seoul metropolitan area. *Cities*, *27*(5), 358–368.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological bulletin*, *95*(2), 334.

Stover, V. W., & Bae, C.-H. C. (2011). Impact of gasoline prices on transit ridership in washington state. *Transportation Research Record*, *2217*(1), 11-18. Retrieved from https://doi.org/10.3141/2217-02 doi: 10.3141/2217-02

Syed, S. I., & Khan, A. M. (2000). Factor analysis for the study of determinants of public transit ridership. *Journal of Public Transportation*, *3*(3), 1.

Taylor, B. D., & Fink, C. (2013). Explaining transit ridership: What has the evidence shown? *Transportation Letters*, *5*(1), 15–26.

Taylor, B. D., Miller, D., Iseki, H., & Fink, C. (2003). Analyzing the determinants of transit ridership using a two-stage least squares regression on a national sample of urbanized areas.

Thompson, C. G., Kim, R. S., Aloe, A. M., & Becker, B. J. (2017). Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. *Basic and Applied Social Psychology*, *39*(2), 81–90.

Thompson, G., Brown, J., & Bhattacharya, T. (2012). What really matters for increasing transit ridership: Understanding the determinants of transit ridership demand in broward county, florida. *Urban Studies*, *49*(15), 3327–3345.

Tiefelsdorf, M., Griffith, D. A., & Boots, B. (1999). A variance-stabilizing coding scheme for spatial link matrices. *Environment and planning A*, *31*(1), 165–180.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, *46*(sup1), 234–240.

Tweede Kamer. (2021). *Fracties.* https://www.tweedekamer.nl/kamerleden_en_commissies/fracties. ((Accessed on 04/30/2021))

Upchurch, C., & Kuby, M. (2014). Evaluating light rail sketch planning: actual versus predicted station boardings in phoenix. *Transportation*, *41*(1), 173–192.

van der Blij, F., Veger, J., & Slebos, C. (2010). Hov op loopafstand het invloedsgebied van hov-haltes. In *Colloquium vervoersplanologisch speurwerk* (pp. 1–15).

Van der Groot, D. (2012). *Ov-reizigers, mobiliteitsstandpunten en politieke keuze.* `https://panteia.nl/uploads/sites/2/2016/12/Rapport-OV-reizigers-2012.pdf`. ((Accessed on 04/30/2021))

van Oort, N., van der Bijl, R., & Verhoof, F. (2017). *European transport conference.* Retrieved from `https://nielsvanoort.weblog.tudelft.nl/the-wider-benefits-of-high-quality-of-public-transport-for-cities/`

Zhao, F., Chow, L., Li, M., & Liu, X. (2005). A transit ridership model based on geographically weighted regression and service quality variables. *Lehman Center for Transportation Research, Florida International University, Miami, Florida. http://lctr.eng.fiu.edu/re-projectlink/finalDO97591_BW.pdf (accessed December 12, 2010)*.

Zhou, M., Wang, D., Li, Q., Yue, Y., Tu, W., & Cao, R. (2017). Impacts of weather on public transport ridership: Results from mining data from different sources. *Transportation research part C: emerging technologies*, *75*, 17–29.

Zuniga-Garcia, N., & Machemehl, R. (2020). Dockless electric scooters and transit use in an urban/university environment. In *99th annual meeting of the transportation research board, washington, dc.*

Zwiers, M. (2018). *De meeste buurten veranderen niet snel van karakter.* Retrieved from `https://www.gebiedsontwikkeling.nu/artikelen/de-meeste-buurten-veranderen-niet-snel-van-karakter/`

9 Appendix

9.1 Omitted bias example

Let's say there are two bus stops, one bus stop has 100 daily boarders, a frequency of two busses per hour and the address density around the stop is 1,000. A second bus stop is on the campus of a university and has 500 daily boarders, a frequency of three busses per hour and an address density of 1,000. If a model is estimated to explain bus ridership is based on only the variables frequency and address density, the model would think the frequency is of a major importance for the number of boarders, while the actual increase was caused by the presence of the university. If the presence of the university was included in the model, the model would estimate the importance of frequency more accurate.

9.2 Verbindingswijzer

The Verbindingswijzer (VBW) is an unique tool owned by Movares and is useful to retrieve insights in accessibility and in travel times. All kinds of transport modes can be analysed and many parameters can be adjusted, such as the maximum amount of transfers, the access or egress time and the total travel time. Furthermore, the tool allows to plan in new roads or public transportation routes, after which the effect of the new road or the new route on the accessibility can be analysed easily. For this study, the tool is used to retrieve values for the accessibility variables for residents, jobs, hospitals, higher education, secondary education and P+R. Furthermore, the tool is used to calculate the travel times for cycling and taking the bus between bus stop locations to the nearest train station.

9.3 Calculating demographics per bus stop

Every bus stop has a catchment area with a radius of 400 meters. How the demographics variables are calculated for these areas will be explained by using an example of how the percentage of youth is calculated. In figure 19 there are multiple catchment areas visible, but the example will be about the bus stop in the middle, bus stop Grotestraat in Drees. The catchment area contains 4 zip codes, there are two zip code areas that are not adjacent but are the same zip code. The zip codes have 10, 25, 70 and a unknown number of youth and 80, 210, 765 and 75 residents. The catchment area exists of 72% of the first zip code, 5% of the second, 21% of the third and 2 percent of the last. The number of youths and residents per zip code is multiplied by the corresponding percentages and summed, resulting in 230 residents and 23 youth in the catchment area, thus the percentage youth is 10%. The maximum value for the unknown number of youth is 4, which would lead to the percentage of youth being 11.7%, a negligible difference.

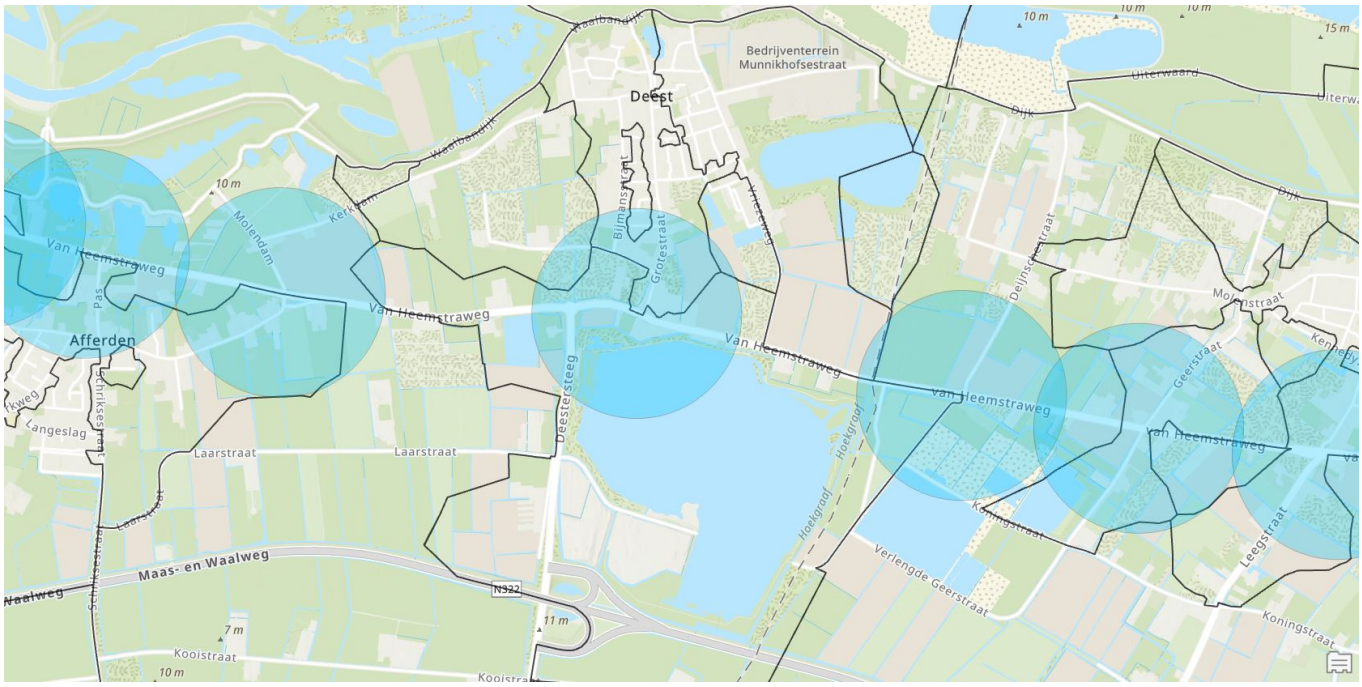

Figure 19: Bus stop Grotestraat in Drees (Screenshot in ArcGIS)

9.4 Box-cox transformation

The Box-Cox transformation uses two lambdas (lambda and lambda2), which differ for each data set, to transform the data as close as possible to a normal distribution (Box & Cox, 1964) This means, that for example the value 10 can be transformed to 0.7 in data set A and to 1.3 in data set B. The equation is displayed below (Box & Cox, 1964)

$$y_{bc} = (((y + \lambda_2)^{\lambda}) - 1)/\lambda) \tag{6}$$

To reverse the transformation, following equation can be applied.

$$y = \lambda_2 + (((y_{bc} * \lambda) + 1)^{(}1/\lambda) \tag{7}$$

9.5 Other bus stops within catchment area

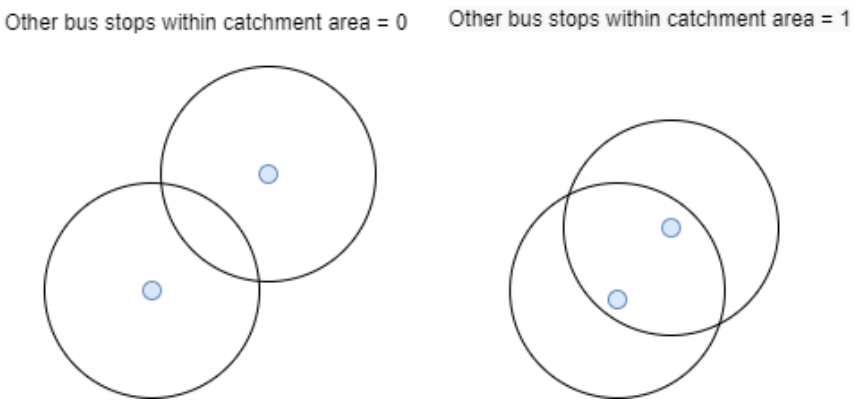

Figure 20: Other bus stops within catchment area

Table 19: Skewness

	N	Skewness	
	Statistic	Statistic	Std. Error
Rides	965	17.753	0.079
Residents	965	0.889	0.079
Average household size	965	1.357	0.079
Average car ownership	965	-0.634	0.079
Female	965	-2.479	0.079
Youth (15-24)	965	3.153	0.079
Elderly (65+)	965	0.422	0.079
Households with low income	965	0.522	0.079
Western background	965	-1.086	0.079
Non Western background	965	1.875	0.079
Green votes	965	0.256	0.079
LU residential	965	-0.213	0.079
LU business	965	6.156	0.079
LU social and commercial	965	4.460	0.079
LU agricultural	965	6.544	0.079
LU recreational	965	2.863	0.079
LU amusement	965	10.816	0.079
Address density	965	2.662	0.079
Distance to station	965	1.709	0.079
Parking rate	965	9.531	0.079
Jobs within buffer	965	7.263	0.079
Train stations within buffer	965	3.958	0.079
Hospitals within buffer	965	8.133	0.079
Higher education within buffer	965	9.220	0.079
Secondary education within buffer	965	3.154	0.079
P+R within buffer	965	5.978	0.079
Other bus stops within buffer	965	1.298	0.079
Closest to train station	965	6.253	0.079
Closest to hospital	965	13.806	0.079
Closest to higher education	965	15.460	0.079
Closest to secondary education	965	5.273	0.079
Closest to P+R	965	10.225	0.079
Residents accessible	965	1.728	0.079
Jobs accessible	965	0.132	0.079
Hospitals accessible	965	0.454	0.079
Higher education accessible	965	0.532	0.079
Secondary education accessible	965	1.198	0.079
P+R accessible	965	0.765	0.079
Frequency	965	0.474	0.079
Directions	965	6.566	0.079
Shelter	965	-0.558	0.079
Bench	965	-0.687	0.079
Digital information display	965	1.108	0.079
Bicycle stand	965	1.012	0.079
Accessible for disabled	965	-0.888	0.079

9.6 Operationalisation

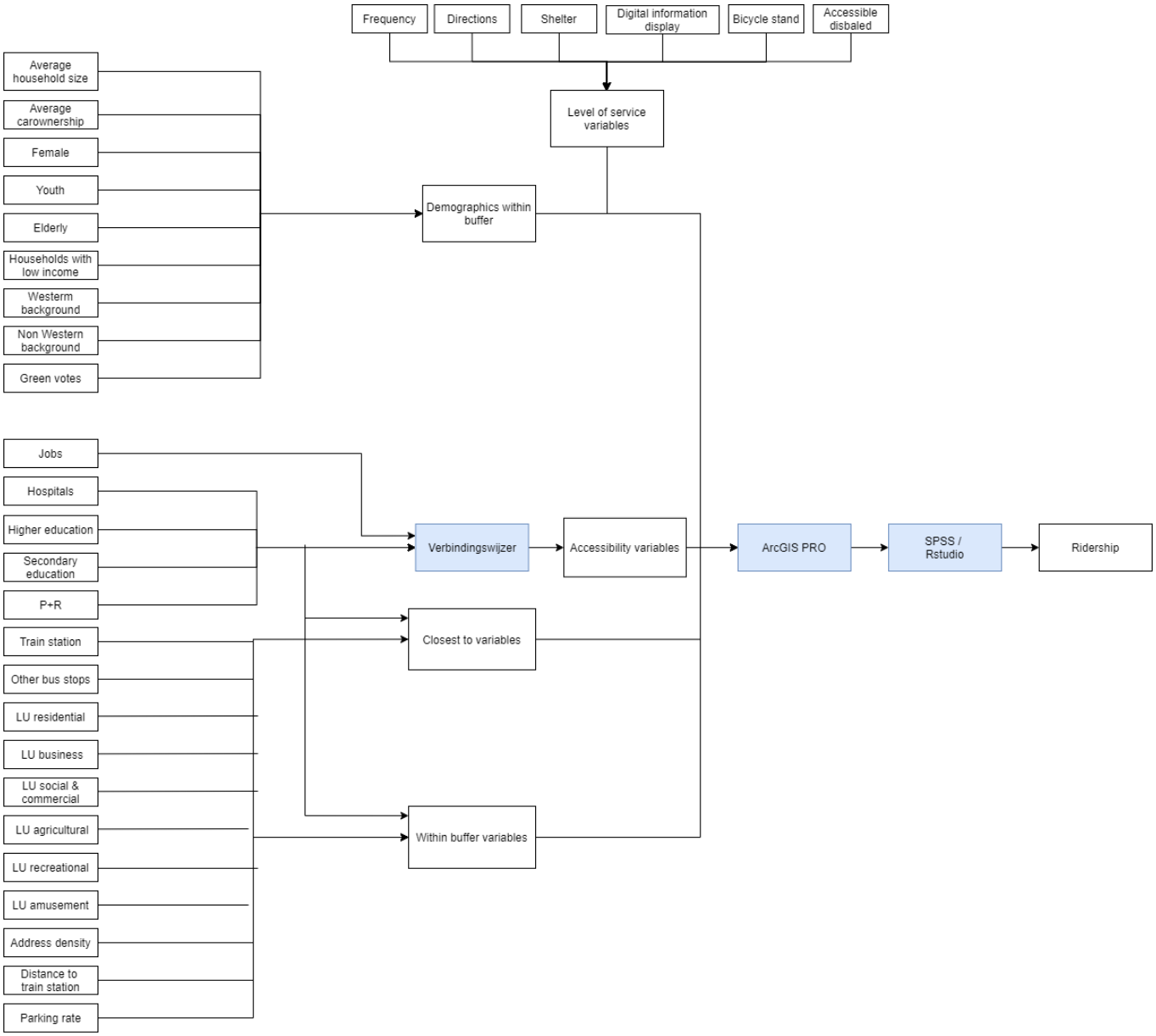

Figure 21: Conceptualisation opreationalisation

9.7 VIF

Table 20: VIF

	Unstandardized Coefficients		Standardized Coefficients				
	B	Std. Error	Beta	t	Sig.	Tolerance	VIF
(Constant)	2.69111	2.938261		0.915885	0.359967		
Residents	-9.3E-05	0.00022	-0.02995	-0.42162	0.673403	0.055104	18.14753
Average household size	0.16602	0.316784	0.017349	0.524078	0.60035	0.253821	3.93978
Average car ownership	-3.30169	1.710346	-0.07543	-1.93042	0.053861	0.182186	5.488885
Female	0.018706	0.036717	0.011848	0.509448	0.610561	0.51425	1.94458
Youth (15-24)	0.045342	0.018273	0.071265	2.481446	0.013263	0.337236	2.96528
Elderly (65+)	-0.02481	0.011931	-0.05554	-2.07978	0.037822	0.389981	2.564231
Households with low income	0.0049	0.009886	0.016346	0.495624	0.620278	0.255721	3.910513
Western background	-0.00672	0.030966	-0.00437	-0.21691	0.828326	0.686155	1.457397
Non Western background	0.004646	0.00995	0.013292	0.466959	0.64064	0.343302	2.912891
Green votes	0.008591	0.007938	0.030408	1.082331	0.279389	0.352379	2.837857
LU residential	0.01788	0.004547	0.140427	3.932079	9.05E-05	0.218081	4.585453
LU business	-0.00604	0.007393	-0.01497	-0.81639	0.41449	0.82767	1.208211
LU social and commercial	0.002516	0.01315	0.004395	0.191314	0.848322	0.526981	1.8976
LU agricultural	-0.00305	0.005855	-0.00913	-0.52154	0.602117	0.907835	1.101522
LU recreational	0.008875	0.007808	0.021205	1.136705	0.255958	0.799275	1.251134
LU amusement	0.012743	0.019973	0.011407	0.638019	0.52362	0.870105	1.149287
Address density	-2.2E-06	3.17E-05	-0.00408	-0.07102	0.943401	0.084429	11.84433
Distance to station	0.043303	0.01299	0.076188	3.333502	0.000892	0.532486	1.877982
Parking rate	21.29664	20.99877	0.017931	1.014185	0.310761	0.889795	1.123855
Jobs within buffer	-7.1E-05	8.4E-05	-0.02135	-0.85008	0.3955	0.440801	2.268599
Train stations within buffer	-0.02752	0.386665	-0.00176	-0.07117	0.943278	0.455688	2.194484
Hospitals within buffer	0.678771	0.646675	0.022967	1.049632	0.294163	0.580938	1.721354
Higher education within buffer	2.693259	0.784934	0.080906	3.43119	0.000628	0.500269	1.998926
Secondary education within buffer	0.010606	0.173212	0.001385	0.061229	0.95119	0.543758	1.839053
P+R within buffer	-0.90506	0.514981	-0.04069	-1.75746	0.079172	0.518992	1.92681
Other bus stops within buffer	-0.06858	0.061917	-0.02176	-1.10759	0.268327	0.720922	1.387112
Closest to train station	0.829981	0.556433	0.035825	1.49161	0.136144	0.482177	2.073926
Closest to hospital	1.463906	1.033693	0.029742	1.41619	0.157058	0.630647	1.585672
Closest to higher education	-1.71771	1.193348	-0.03123	-1.4394	0.150376	0.590871	1.692415
Closest to secondary education	0.327519	0.271092	0.025906	1.208147	0.227301	0.604952	1.653025
Closest to P+R	1.159276	0.809969	0.031533	1.431259	0.152696	0.573025	1.745126
Residents accessible	-4.2E-06	4.08E-06	-0.06783	-1.042	0.297684	0.065631	15.23669
Jobs accessible	1.73E-05	5.91E-06	0.117088	2.934522	0.003424	0.174714	5.723652
Hospitals accessible	-0.13349	0.113611	-0.03041	-1.17494	0.240324	0.415098	2.409068
Higher education accessible	0.077378	0.104614	0.018235	0.739653	0.459699	0.457632	2.18516
Secondary education accessible	0.032909	0.027231	0.06545	1.208501	0.227165	0.09483	10.54514
P+R accessible	-0.25421	0.100276	-0.09879	-2.53507	0.011407	0.18316	5.45972
Frequency	1.727339	0.095458	0.510794	18.09533	7.62E-63	0.349074	2.864725
Directions	0.025224	0.028967	0.020956	0.87078	0.384101	0.480274	2.082145
Shelter	0.256498	0.215511	0.034961	1.190189	0.234279	0.322365	3.102077
Bench	0.737392	0.220361	0.098694	3.346289	0.000852	0.319759	3.12736
Digital information display	1.249793	0.1633	0.154748	7.653355	4.94E-14	0.68034	1.469853
Bicycle stand	0.073099	0.157398	0.009232	0.464422	0.642456	0.70383	1.420798
Accessible for disabled	0.429135	0.165265	0.055511	2.596645	0.009564	0.608605	1.643101

Table 21: First model

	Unstandardized Coefficients		Standardized Coefficients		Sig.	Collinearity Statistics	
	B	Std. Error	Beta	t		Tolerance	VIF
(Constant)	2.568546	2.934405		0.875321	0.381627		
Average household size	0.170576	0.316605	0.017825	0.538765	0.590179	0.253898	3.938593
Average car ownership	-3.15109	1.691295	-0.07199	-1.86312	0.062763	0.186159	5.37175
Female	0.018135	0.03621	0.011487	0.500832	0.616609	0.528319	1.892796
Youth (15-24)	0.043904	0.017643	0.069005	2.488447	0.013006	0.361423	2.766844
Elderly (65+)	-0.02335	0.011063	-0.05226	-2.11012	0.035117	0.453176	2.206646
Households with low income	0.005745	0.009833	0.019167	0.584294	0.559166	0.258261	3.872047
Western background	-0.00816	0.030883	-0.0053	-0.26408	0.79178	0.689286	1.450777
Non Western background	0.003822	0.009877	0.010933	0.386941	0.698889	0.348123	2.872551
Green votes	0.009118	0.007913	0.032273	1.152225	0.249527	0.354253	2.822841
LU residential	0.016407	0.003445	0.128855	4.761761	2.23E-06	0.379533	2.634818
LU business	-0.00666	0.007368	-0.01651	-0.90345	0.366524	0.832446	1.201279
LU social and commercial	0.001042	0.013078	0.00182	0.079653	0.936531	0.532429	1.878186
LU agricultural	-0.00308	0.005852	-0.0092	-0.52559	0.599303	0.907883	1.101463
LU recreational	0.008804	0.007795	0.021036	1.129419	0.259015	0.80115	1.248205
LU amusement	0.014677	0.019876	0.013138	0.738409	0.460454	0.877891	1.139094
Address density	-1.6E-05	1.86E-05	-0.02905	-0.86275	0.388498	0.245192	4.07843
Distance to station	0.043306	0.012662	0.076192	3.420259	0.000653	0.560026	1.785631
Parking rate	20.48021	20.97637	0.017244	0.976347	0.329149	0.890956	1.122389
Jobs within buffer	-8.6E-05	8.24E-05	-0.02563	-1.03945	0.29887	0.457009	2.188141
Train stations within buffer	-0.07969	0.383382	-0.00509	-0.20787	0.835377	0.46314	2.159173
Hospitals within buffer	0.662055	0.644708	0.022402	1.026907	0.304734	0.584004	1.712318
Higher education within buffer	2.686449	0.783561	0.080701	3.428514	0.000634	0.501607	1.993591
Secondary education within buffer	0.021987	0.172509	0.002871	0.127453	0.89861	0.547745	1.825668
P+R within buffer	-0.9019	0.514673	-0.04054	-1.75237	0.080042	0.519182	1.926108
Other bus stops within buffer	-0.0695	0.061708	-0.02205	-1.12631	0.260329	0.725217	1.378898
Closest to train station	0.853632	0.555787	0.036846	1.535899	0.124906	0.482898	2.070831
Closest to hospital	1.441	1.032912	0.029276	1.395085	0.163326	0.631076	1.584594
Closest to higher education	-1.79507	1.190731	-0.03264	-1.50754	0.132015	0.59298	1.686399
Closest to secondary education	0.322925	0.27094	0.025542	1.19187	0.233619	0.605129	1.65254
Closest to P+R	1.193879	0.809026	0.032475	1.475699	0.140366	0.573885	1.74251
Jobs accessible	1.42E-05	5.04E-06	0.095666	2.809779	0.005062	0.239743	4.171129
Hospitals accessible	-0.16203	0.108596	-0.03692	-1.49206	0.136025	0.453947	2.2029
Higher education accessible	0.07484	0.104459	0.017637	0.716458	0.47389	0.458614	2.180485
Secondary education accessible	0.01684	0.02233	0.033491	0.75412	0.450969	0.140906	7.096951
P+R accessible	-0.25503	0.099921	-0.09911	-2.55233	0.010861	0.184312	5.425582
Frequency	1.716717	0.094895	0.507653	18.09064	7.78E-63	0.35293	2.833422
Directions	0.021983	0.028805	0.018263	0.763181	0.44555	0.485297	2.060595
Shelter	0.268184	0.215015	0.036553	1.247277	0.212613	0.323583	3.0904
Bench	0.731209	0.220197	0.097866	3.320704	0.000933	0.31997	3.125295
Digital information display	1.258348	0.162886	0.155808	7.725323	2.91E-14	0.683234	1.463627
Bicycle stand	0.069436	0.157248	0.00877	0.44157	0.658904	0.704589	1.419267
Accessible for disabled	0.436613	0.164106	0.056479	2.660549	0.007937	0.616719	1.621485

9.8 Scatterplot difference travel time

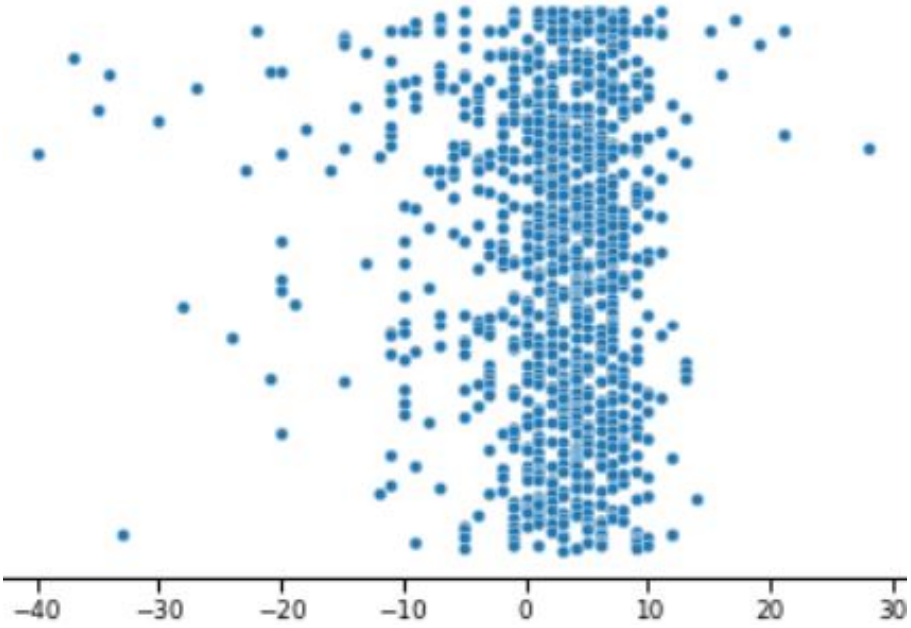

Figure 22: Scatterplot difference travel time

9.9 Full OLS model excluding frequency

Table 22: Full model, excluding frequency

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	2.558	3.936		0.650	0.516
Average household size	-0.039	0.425	-0.004	-0.091	0.927
Average car ownership	-3.154	2.270	-0.072	-1.389	0.165
Female	0.027	0.049	0.017	0.553	0.580
Youth (15-24)	0.063	0.024	0.100	2.667	0.008
Elderly (65+)	-0.046	0.015	-0.103	-3.069	0.002
Households with low income	0.031	0.013	0.103	2.355	0.019
Western background	0.023	0.042	0.015	0.534	0.593
Non Western background	0.004	0.013	0.011	0.280	0.780
Green votes	0.018	0.010	0.064	1.723	0.085
LU residential	0.024	0.005	0.190	5.311	0.000
LU business	-0.016	0.010	-0.039	-1.592	0.112
LU social and commercial	0.033	0.018	0.056	1.835	0.067
LU agricultural	-0.012	0.008	-0.036	-1.542	0.123
LU recreational	0.019	0.010	0.045	1.780	0.075
LU amusement	0.045	0.027	0.040	1.679	0.093
Address density	-4.148E-05	0.000	-0.075	-1.665	0.096
Distance to station	0.069	0.017	0.122	4.100	0.000
Parking rate	74.864	28.123	0.063	2.662	0.008
Jobs within buffer	0.000	0.000	-0.087	-2.591	0.010
Train stations within buffer	-0.928	0.514	-0.059	-1.806	0.071
Hospitals within buffer	1.717	0.869	0.058	1.976	0.048
Higher education within buffer	0.468	1.044	0.014	0.448	0.654
Secondary education within buffer	0.319	0.232	0.042	1.372	0.170
P+R within buffer	-0.795	0.694	-0.036	-1.146	0.252
Other bus stops within buffer	-0.402	0.081	-0.128	-4.948	0.000
Closest to train station	2.534	0.747	0.110	3.391	0.001
Closest to hospital	1.671	1.392	0.034	1.201	0.230
Closest to higher education	0.386	1.653	0.007	0.233	0.815
Closest to secondary education	-0.055	0.365	-0.004	-0.150	0.881
Closest to P+R	1.656	1.093	0.045	1.516	0.130
Jobs accessible	4.215E-05	0.000	0.285	6.303	0.000
Hospitals accessible	-0.248	0.145	-0.057	-1.708	0.088
Higher education accessible	0.384	0.140	0.091	2.742	0.006
Secondary education accessible	0.025	0.030	0.051	0.844	0.399
P+R accessible	-0.276	0.135	-0.108	-2.048	0.041
Dummy central station	-6.779	2.275	-0.087	-2.979	0.003
Directions	0.435	0.043	0.362	10.055	0.000
Bus operators	-0.360	0.219	-0.048	-1.640	0.101
Directions	0.025	0.029	0.021	0.871	0.384

9.10 Assumptions OLS

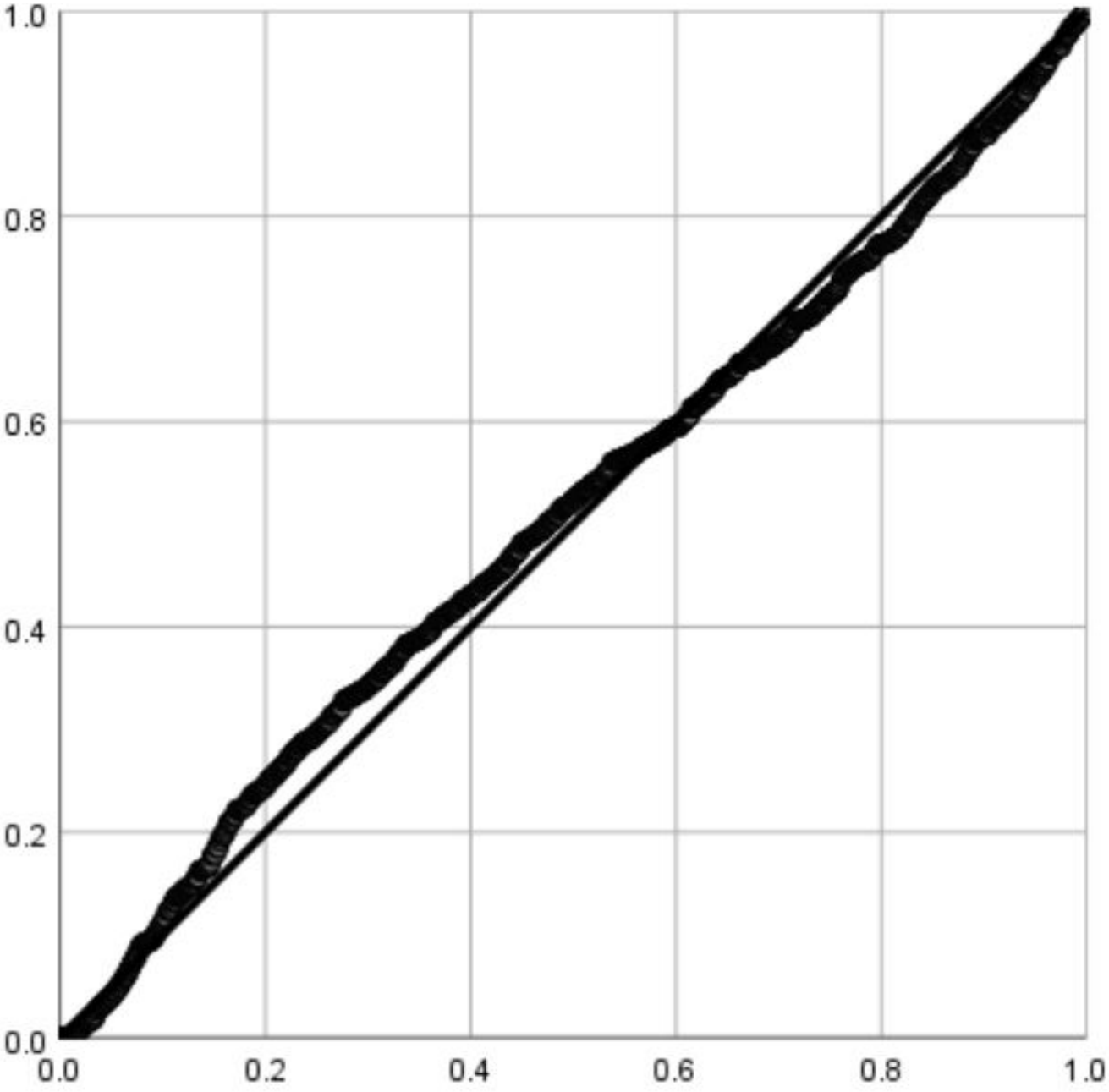

Figure 23: P-P plot

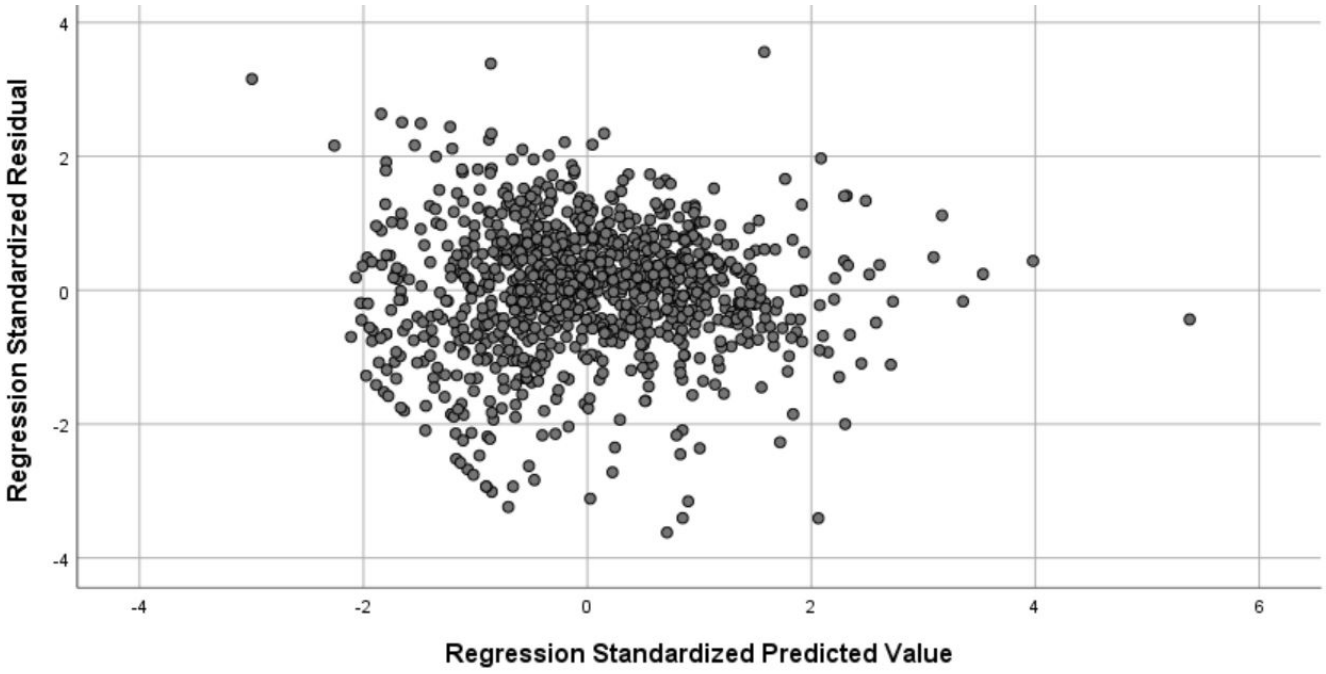

Figure 24: Variance residuals

9.11 Correlation car ownership and low income households

Table 23: Correlation car ownership and low income households

		Average car ownership	Percentage low income households
Average car ownership	Pearson Correlation	1	-.751**
	Sig. (2-tailed)		0.000
	N	964	964
Percentage low income households	Pearson Correlation	-.751**	1
	Sig. (2-tailed)	0.000	
	N	964	964

9.12 Correlations frequency and accessibility higher education and hospitals

Table 24: Correlations frequency and accessibility higher education and hospitals

		Frequency	Accessibility higher education	Accessibility hospitals
Frequency	Pearson Correlation	1	.379**	.296**
	Sig. (2-tailed)		0.000	0.000
	N	964	964	964
Accessibility higher education	Pearson Correlation	.379**	1	.368**
	Sig. (2-tailed)	0.000		0.000
	N	964	964	964
Accessibility hospitals	Pearson Correlation	.296**	.368**	1
	Sig. (2-tailed)	0.000	0.000	
	N	964	964	964

9.13 Secondary schools

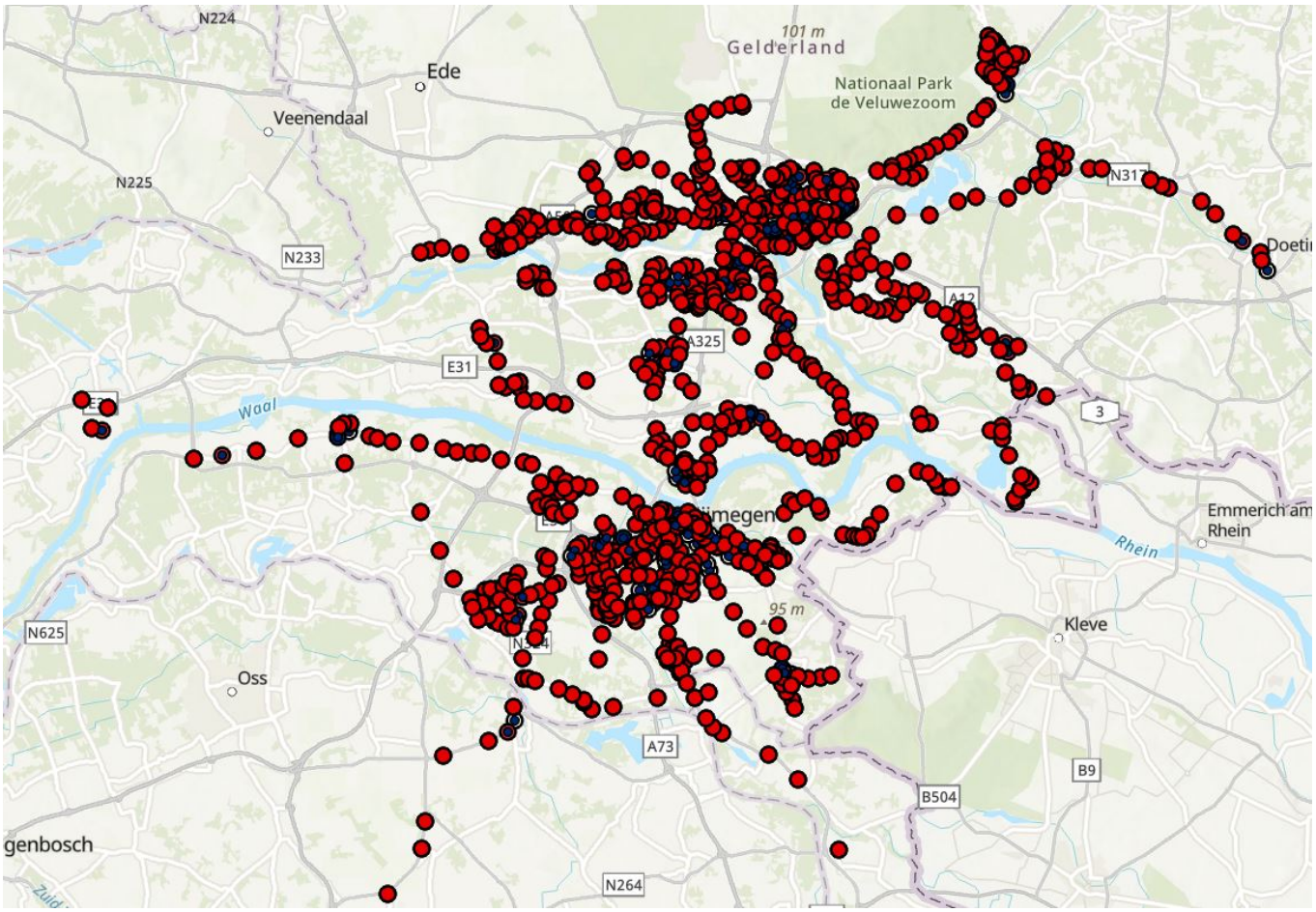

Figure 25: Bus stops with a secondary school in catchment area

9.14 Car ownership

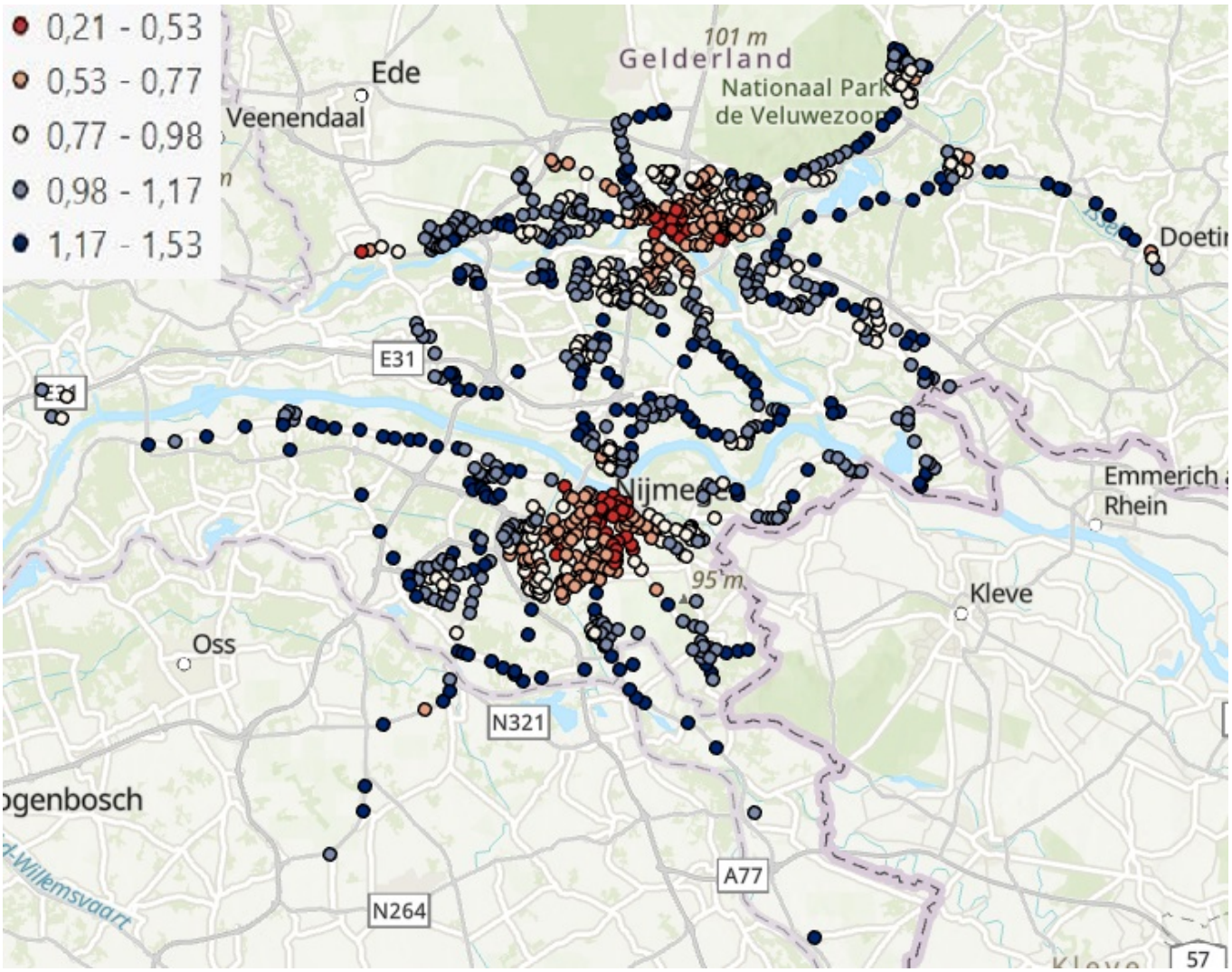

Figure 26: Car ownership

9.15 Access time VBW

It is possible that the bus stop locations in the VBW and the bus stop locations used in this study did not exactly match. A maximum walking time of ten minutes was set in the VBW to ensure the bus stop could be reached. When for example, the maximum time was set to one minute and the bus stop location in the VBW and the bus stop location used in this study were two minutes walking apart. Therefore no bus stop can be reached and the values for the accessibility variables would be very low.

9.16 Estimates SLX model

Table 25: SLX full model

	B	Std. Error	Std. B	Sig.
(Intercept)	-0.689	3.131		0.826
Average household size	0.545	0.313	0.056	0.082*
Average car ownership	-1.064	0.698	-0.075	0.128
Female	-0.009	0.050	-0.005	0.861
Youth (15-24)	0.041	0.019	0.063	0.032**
Elderly (65+)	-0.009	0.017	-0.021	0.589
Households with low income	0.012	0.011	0.038	0.278
Western background	0.031	0.048	0.019	0.523
Non Western background	-0.004	0.010	-0.012	0.689
Green votes	0.002	0.009	0.008	0.796
LU residential	0.037	0.005	0.284	0.000***
LU business	-0.004	0.009	-0.010	0.659
LU social and commercial	0.068	0.017	0.114	0.000***
LU agricultural	-0.001	0.007	-0.003	0.902
LU recreational	0.022	0.010	0.051	0.036**
LU amusement	0.036	0.031	0.032	0.250
Address density	-1.17E-4	4.19E-5	-0.014	0.709
Distance to station	0.007	0.026	0.012	0.800
Parking rate	14.311	22.651	0.012	0.528
Jobs within buffer	-4E-5	-1E-4	-0.012	0.737
Hospitals within buffer	1.660	0.744	0.055	0.026**
Higher education within buffer	1.539	0.921	0.045	0.095*
Secondary education within buffer	0.191	0.176	0.024	0.280
P+R within buffer	-0.527	0.561	-0.023	0.348
Other bus stops within buffer	-0.052	0.073	-0.016	0.472
Closest to train station	1.413	0.471	0.060	0.003***
Jobs accessible	-7.69E-6	1.04E-5	0.118	0.002***
Hospitals accessible	-0.359	0.199	-0.080	0.071*
Higher education accessible	-0.117	0.174	-0.027	0.501
Secondary education accessible	-0.018	0.025	-0.034	0.477
P+R accessible	-0.050	0.110	-0.019	0.647
Dummy central station	3.673	1.813	0.046	0.043
Frequency	2.602	0.118	0.636	0.000***
Directions	0.103	0.041	0.084	0.011**
Bus operators	-0.613	0.213	-0.08	0.004***
Lag x Female	0.104	0.066	0.049	0.116
Lag x Elderly (65+)	-0.036	0.020	-0.069	0.074*
Lag x Western background	0.020	0.062	0.010	0.745
Lag x LU residential	-0.028	0.006	-0.168	0.000***
Lag x LU business	-0.005	0.013	-0.009	0.693
Lag x LU social and commercial	-0.048	0.024	-0.068	0.044**
Lag x LU agricultural	-0.007	0.012	-0.013	0.573
Lag x LU recreational	-0.003	0.015	-0.005	0.838
Lag x LU amusement	0.010	0.042	0.007	0.812
Lag x Distance to station	0.056	0.028	0.088	0.044**
Lag x Parking rate	-33.025	39.797	-0.017	0.407
Lag x Jobs within buffer	-1.88E-4	1.77E-04	-0.004	0.920

Table 25 continued from previous page

	B	Std. Error	Std. B	Sig.
Lag x Hospitals within buffer	0.546	1.079	0.013	0.613
Lag x Higher education within buffer	0.002	1.219	0.000	0.998
Lag x Secondary education within buffer	0.351	0.243	0.036	0.149
Lag x P+R within buffer	0.589	0.743	0.020	0.428
Lag x Other bus stops within buffer	-0.233	0.102	-0.058	0.023**
Lag x Closest to train station	-0.480	0.889	-0.011	0.590
Lag x Hospitals accessible	0.206	0.221	0.043	0.350
Lag x Higher education accessible	0.322	0.199	0.067	0.106
Lag x Dummy central station	9.355	3.186	0.079	0.003***
Lag x Frequency	-0.171	0.160	-0.034	0.286
Lag x Directions	-0.189	0.062	-0.116	0.002***
Lag x Bus operators	0.175	0.291	0.017	0.548