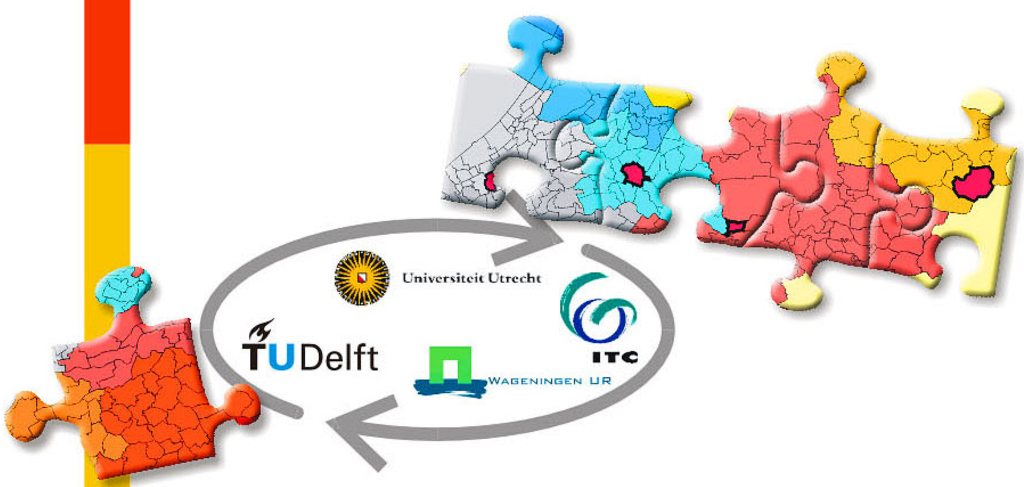
Towards a GIS based method for ridership forecasting for newly planned train stations in The Netherlands

MSc thesis

Author: Sjimmie Dirkx

Supervisor: Fred Toppen (UU)
Professor: Stan Geertman (UU)



Towards a GIS based method for ridership forecasting for newly planned train stations in The Netherlands

MSc thesis
February 2012

Author: Sjimmie Dirkx

Supervisor: Fred Toppen (UU)
Professor: Stan Geertman (UU)

Abstract

Between 2004 and 2010, 22 new train stations were opened in The Netherlands. In 2011, again six new stations were opened. These include Emmen Zuid, Veendam, and Giessendam Blauwe Zoom. More than a dozen are planned to be opened in 2012. The list of stations to be opened in the future includes over 70 concrete and very unspecific plans for other new stations. Before a station is opened, forecasts are made regarding expected passenger flows, usually expressed in number of average weekday boardings. It is important that these ridership forecasts are as accurate as possible, since the economic viability of a station depends on the number of passengers. It is often considered a loss of community funds when large demand shortfalls occur. Unfortunately, they frequently do. Three main causes for inaccurate forecasts can be identified. First, it is inaccurate data that affects the outcomes. The principle of garbage in, garbage out has its effects. This should lead to both overly optimistic and overly pessimistic forecasts. These should eventually average out. The truth is, nevertheless, that ridership forecasts tend to be overly optimistic. The over-optimism is generally not averaged out by pessimistic forecasts, because the inaccuracies are caused by two phenomena associated with ridership forecasting, known as optimism bias and strategic misrepresentation. The former refers to the human tendency of planners and forecasters to be positive about the project at hand, since a generally negative attitude and disbelief towards the project would not get it to the forecasting phase in the first place. The second problem that is at the root of overly optimistic forecasts is strategic misrepresentation, which is the deliberate manipulation of forecasts in order to get projects approved. When authorities are competing with each other over higher level funding, they are encouraged to present their project as positive as possible. It will be, in the end, the project that will be most beneficial to walk away with the funding. Though the practice of strategic misrepresentation is difficult to prove, its existence has been confirmed in recent literature.

The goal of the research was therefore to find the building blocks needed to develop a method for ridership forecasting that leaves little to no room for optimism bias and strategic misrepresentation, and thus be as accurate and unbiased as possible. The method is aimed at stop train stations only, since they make up the vast majority of newly opened and proposed new stations. The method is designed in a GIS environment because proximity to a train station is a major aspect in people's likelihood to travel by train. Furthermore, a GIS based method enables quick and relatively easy calculation of forecasts for newly proposed stations, by keeping the calculations within a single environment.

In the development phase, GIS is used to determine the stations' spheres of influence based on the road network, rather than crow-flight distance common for conventional methods. Subsequently, the service areas determined in this way are used to acquire population and environment data for the stations. Together with station data on accessibility and network characteristics, these are used to determine multiple regression functions, which can be used to asses the variables that are significant for ridership in the Dutch case. The regression functions are based on several different reference classes of existing Dutch stations. The resulting forecasts are thus based on historical precedent. The use of reference classes is the proposed way to go in order to rule out optimism bias and strategic misrepresentation. To asses the quality of the forecasts resulting from the different multiple regression functions and their respective reference classes, 17 recently opened train stations are used. This enables the comparison of forecasts for these stations with actual ridership numbers measured in 2009 and 2010.

It is found that the method is indeed unbiased and has thus, successfully ruled out optimism bias. Furthermore, it leaves little room for strategic misrepresentation. The outcomes can only be manipulated at the data acquisition, adaptation, and application phases. To asses the precise quality of the forecasts generated by the proposed method, more time is needed. This is because it usually takes new stations several years to mature and grow to their full potential. It has yet been found though, that the model leaves room for improvement and recommendations for modifications have thus been made.

Contents

1 Introduction

1.1 Background

In 1839, the first railway in The Netherlands was opened between the cities of Haarlem and Amsterdam. The rail network grew and quickly connected several other cities in the west of the country and later almost the whole of the country. There are nowadays over 2800 kilometers of railway in The Netherlands, connecting almost 400 train stations all over the country. Total ridership numbers are over a million, which means every day, over 500,000 people travel by train in The Netherlands. Of a total population of over 16 million, 62 percent makes use of the train at least once a year. All in all, the train is a mode of transportation of significance.

The Dutch railways are split up between an organization responsible for the maintenance and development of the tracks and other infrastructure on the one hand and a number of passenger railway operating companies on the other. The former, named ProRail, is owned by the state. Among the latter are NS, Arriva, Connexxion, Syntus, and Veolia Transport. Of these, NS is by far the largest, operating the core national railway network, connecting the major cities of the country. The other operating companies cover the subsidiary regional lines. Most of these lines are located in the eastern and northern parts of the country, but examples can also be found elsewhere.

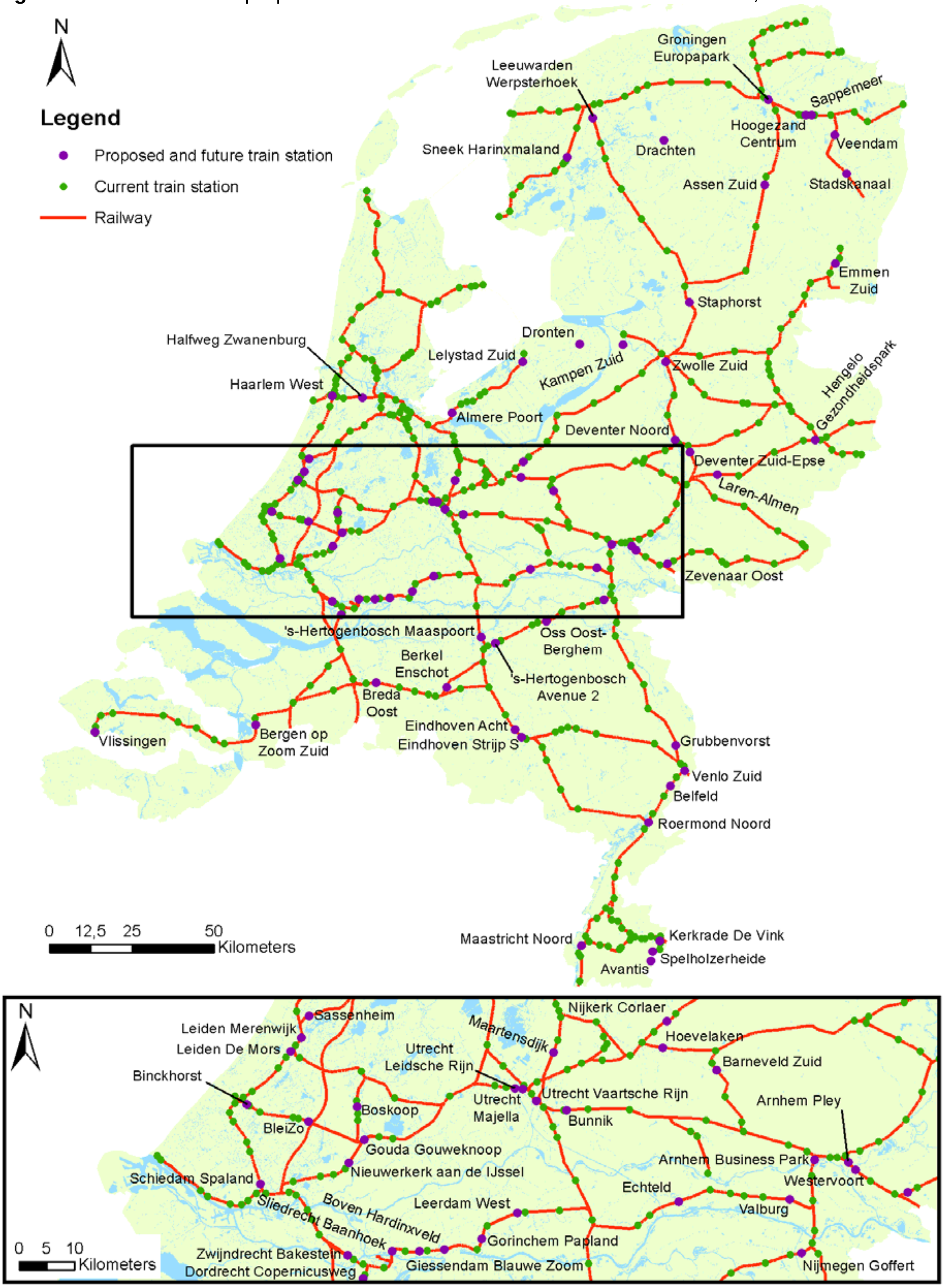
Public transportation in general and the railways specifically are often up for political debate. Because ProRail and NS are both still fully state-owned, the people's representatives often question the Minister on a wide range of subjects regarding passenger railway transportation, such as availability of toilet facilities, quality of schedules, delays and punctuality, and the opening of new railway stations. The latter has resulted in the document "Nieuwe stations, nieuwe passagiers" (New stations, new passengers) by the Dutch House of Representatives (Tweede Kamer der Staten-Generaal, 2008). This document investigates proposed and planned railway stations and the process to the opening of new railway stations. It subsequently provides a list, though incomplete, of existing plans for new stations.

In total, around 70 proposed railway stations can be counted in The Netherlands (see appendix 12 for an incomplete list). Examples are Berkel-Enschot, Boven Hardinxveld, Leerdam West, Spekholzerheide, and Maartensdijk (see figure 1.1). Their statuses are diverse. Some, like Kampen Zuid en Boven Hardinxveld, are certain to be developed within the next few years, while others are no more than part of a rough draft plan. There are examples of stations that will be developed in case of the construction of a new residential area, such as Leerdam West, while others are dependent on the construction of new tracks or the revitalization of old ones. Some proposed stations can hardly be considered to be new because they are part of larger plans that simultaneously contain the closing of another, nearby station. A recent example of the execution of such a plan is the opening of Emmen Zuid station and the shutdown of Emmen Bargeres at the same time. De facto, this has been a crossover of one station to a different location. Plans of similar magnitude exist for Boskoop, Vlissingen, Bunnik, and Maartensdijk, which will be at the cost of Hollandsche Rading. Of some, the proposed location is not known yet. In addition to the proposed stations from the stock-taking of the House of Representatives, plans for new stations remain to emerge in municipal politics.

Whether or not a proposed station is eventually developed is dependent on a number of factors. Of course, the total costs are an important aspect. It is sought after that all costs of construction, incorporation into existing timetables, and maintenance are covered. Another prerequisite for development is a guarantee by the railway operating company that holds the concession that it will service the new station. The operating companies are not necessarily obliged to act accordingly. Consequently, agreements regarding frequency of stops need to be drawn up in advance (Ministerie van Verkeer en Waterstaat, 2009, p.106). Additional stops due to a new station on an existing route will likely cause the total travel time for most passengers to increase. This increase has to be in proportion with the accumulation of new passengers and, consequently, improved ridership.

Ridership forecasts are a major criterion in the decision-making process for proposed railway stations in The Netherlands (Vergouwen & Baggen, 2004, p.8). Therefore, it is important that the expectations do not divert too much from the eventual ridership numbers. When demand shortfalls do occur, especially in the case of extremely large overestimation, the financial viability becomes at stake. Because conventional methods for ridership estimation are notorious for their overestimations of passenger forecasts (Projectteam Toepassing Norm ProRail, 2009) it is justified to look for methods with more accurate estimates.

Figure 1.1: A selection of proposed and future train stations in The Netherlands, 2011



Source: Goudappel Coffeng, 2011; own data, 2011; own adaptation, 2011.

Conventional methods for ridership forecasts are based on crow-flight distance circles around the location for the proposed railway station. Based on the number of residents, the number of jobs, and the number of student places within these circles, calculations are made with given parameters to come up with an expectation. The circles, which would be better described with the term rings since they do not overlap, usually have a 500 meter radius. The number of rings applied depends on the exact method that is chosen. A brief analysis of circle methods applied to several cases by different traffic engineering businesses shows that, though all are based on the same principles, both methods and results can vary significantly (Nanninga & Woolthuis, 2004; Beets & Tertoolen, 2009; Kooij, 2009, pp.21-22).

Research by Gutierrez and Garcia-Palomares (2008) has shown that the use of Euclidean distance buffers leads to overestimation of the coverage area of public transportation stops and consequently also of the number of residents within the station's service area. Others have indicated that transportation demand forecasts are chronically overoptimistic due to misinterpretation of data, lack of knowledge or biased methodology (Tal, 2008; Mackett & Edwards, 1997; Flyvbjerg, 2006; Flyvbjerg, 2007; Flyvbjerg et al., 2005). In order to reduce overestimation and bias in forecasting, the application of a method called reference class forecasting is suggested (Flyvbjerg et al., 2005, pp.140-142). This method is based on taking information from a class of similar cases, based on statistics, in order to establish a forecast with greater certainty than conventional methods that focus mainly on the researched case itself. Reference class forecasting is shown to be more reliable than methods of forecasting not based on statistics and historical precedent.

1.2 Objectives

The main objective of this research is to draw a framework that contains the building blocks required to design and develop a GIS based method for ridership forecasting for newly planned train stations in The Netherlands. The method aims at yielding more accurate and more reliable results than conventional methods used by traffic engineering businesses in The Netherlands. These methods are based on population data derived from crow-flight distance rings around the train station's location. Instead, the new method will be based on service areas derived from network analyses and a number of relevant socio-economic, demographic, and transport aspects rather than plain population numbers and the number of jobs.

Until now, ridership forecasts have usually been presented as round numbers. For example, Assen Zuid station is expected to generate 2000 passengers on a daily basis (Tweede Kamer der Staten-Generaal, 2008, p.24). The research into the new method aims at looking at other options to present ridership forecasts. One should think of an interval within which the eventual actual ridership number will fall with, say, 90 percent certainty. This should lead to less uncertainty about the forecasts, given the fact that no model is suitable to exactly predict future developments and precise numbers therefore appear to be quite suggestive. Expressing the forecasts within an interval is only necessary in case the model shows to be not sophisticated enough to forecast ridership in a very high accuracy.

1.3 Research questions

Since the main objective of this research is to find the requirements for developing a new method for ridership forecasting for newly planned train stations, as described in the previous section, the following main research question has been formulated accordingly:

> *How can a new GIS based method for ridership forecasting for newly planned train stations in The Netherlands be designed in such a way that it will yield more accurate and more reliable results than conventional methods and what should be the model's parameters?*

The main research question is accompanied by a set of sub questions. Different chapters in the thesis will attempt to answer these questions. Subsequently, the answers generated from the sub questions will help to draw the final conclusions. The sub questions are as follows:

1. What are the characteristics of conventional methods for ridership forecasting and which problems are associated with it?
2. How can the overoptimistic forecasts from conventional methods be explained?
3. What are proposed solutions for the problem of overly optimistic forecasts?
4. What are, generally speaking, the features of train passengers and how can the transit dependent traveler be described?
5. Which factors are determinants for passenger railway transportation, apart from passenger characteristics, and consequently play a part in ridership forecasting?
6. What are determinants for ridership from the built environment of train stations and how do they add to the number of daily boardings?

1.4 Scope

The thesis will include recommendations for the aspects that should compose a GIS based method for the calculation of ridership forecasts. Multiple regression functions will be used on a number of recently opened train stations. This will enable comparison of the forecasts with actual ridership data, and comparison with other forecasts. Furthermore, the research will include an inventory of future and planned train stations. Subsequently, ridership forecasts based on the proposed method will be provided.

A distinction is made between nonexistent future train stations, existing train stations opened before 2004, and train stations opened after 2004. Ridership forecasts for existing train stations opened before 2004 are out of scope. Furthermore, the proposed method will not be applicable to intercity train stations, but only to regional and commuter rail stations, commonly referred to as stop train stations. A discussion on the usefulness and validity of the decision-making process for new stations will also not be included, even though the advantages and the requisite of the process are sometimes questioned. It is assumed that process requirements remain untouched or that, in case of adjustments to the terms and conditions, ridership forecasts will remain to be an important aspect of the decision-making process.

1.5 Significance

In section 1.1 it is mentioned that there are various plans for the construction of new train station in The Netherlands. A look at the statuses of these plans shows that ten stations are certain to become operational within the following years, though recent history has proven that the expected dates of opening are often missed as official openings are postponed. In addition to the ten stations that are to be expected, several dozens more are, with significantly less certainty, intended or envisaged as part of spatial or traffic plans. This indicates a great demand for new train stations on a local or regional level, which is affirmed and acknowledged by national politics (Tweede Kamer der Staten Generaal, 2008, p.4). The realization of the intended stations is nonetheless dependent on the number of expected new train passengers generated by the station. The limitation regarding new passengers of at least 1000 has been mentioned. More generically, the aspiration is that the new station will optimally contribute to accessibility against minimal costs (Vergouwen & Baggen, p.8).

It is because of minimum ridership requirements that decent forecasts are a necessity. Sufficient ridership forecasts are required for continuation of the decision-making process of a new station. A reliable method with accurate results is consequently required. Because of overestimation by conventional methods, an attempt to develop a new method is justified.

1.6 The use of GIS

In 1970, Tobler presented his first law of geography, which stated that all things are related, but near things are more related than distant things (Miller, 2004; Tobler, 2004). The circle-based method is based on the idea that near things are more related than distant things. The closer a buffer is to the station, the higher the percentage of the residents that is considered to be traveling by train. The idea that everything is related to everything else is neglected however, since only the number of residents, the number of jobs, and college enrollment are considered to affect train ridership. The method proposed by this research will be based on both aspects of Tobler’s first law. A range of identified factors associated with train ridership will be tested for significance (everything is related to everything else) and it is assumed that proximity is the determinant for the extent to which it affects a station’s ridership number.

Nearness and proximity, in Todler’s law, are not necessarily related to Euclidean distance. Distance in units of time, travel costs, and network distance can also be applied. As such, the principle of proximity can take on different meanings in different situations (Tobler, 2004, p.306). The different concepts of distance can cause ambivalence. However, GIS provide the means to apply the different concepts of distance. Effects of phenomena such as congestion, terrain, land cover, infrastructure, and traffic control can be captured by the use of least-cost path tools and network analyses in GIS.

The new method will be designed in ArcMap, which is a mainstream GIS environment. The use of GIS enables one to combine statistical and geographical data. This is especially useful in the case of ridership forecasting, since statistics on several socio-economic and demographic factors and proximity to the train station are said to have great effect on a station's ridership. Furthermore, ArcGIS provides options for the creation of models, which can be used for the proposed method for ridership forecasting.

1.7 Contents

The thesis report is built up logically. The introduction is followed by the theoretical chapters. The theory will lead to a conceptual model. The theoretical part is split up into two chapters, covering ridership forecasting and the problems associated with it in chapter 2 on the one hand and drivers of ridership in chapter 3 on the other hand. The theoretical part is succeeded by a methodological chapter. Chapter 4 discusses the procedures, calculations, calibrations, and techniques used during the research. The choices that are made are substantiated in this chapter as well. Furthermore, limitations of the methodology are mentioned and discussed. References are made to specific software that is applied. The results are presented in chapter 5. This includes the results of statistical analysis with multiple regression which will show the level of significance of the various independent variables on ridership. This chapter is succeeded by the discussion in chapter 6. Here, the results presented in previous chapter are discussed. Presentations and discussion of the results are separated to make a clear distinction between the results in and of themselves and the interpretations of the results. The chapter will discuss patterns, relationships, and trends in the results. Whether the results are in compliance with findings from theoretical chapters are covered. Subsequently, the significance of the results will be discussed. The final chapter (7) of the thesis presents the conclusions and limitations of the research as well as recommendations for further research to fill in remaining blanks or for improvement.

2 Ridership forecasting

2.1 Introduction

Exploring the future and making forecasts of future events are important activities for both commercial and governmental organizations. They are focused on the generation of specific intelligence. As such, the outcomes are used as a base for new policies (Stuurgroep Toekomstonderzoek en Strategisch Omgevingsbeleid, 2000, p.79). In the case of the proposed construction of a new train station, ridership forecasts, together with budgetary estimates, are used to determine the viability of the proposed station. Accurate, reliable results are thus important, since incorrect outcomes may lead to the construction of economically unviable train stations due to cost overruns and demand shortfalls. Overoptimistic forecasts therefore pose a threat to the economy and leave governments with financial problems. This will eventually affect local tax payers. Based on findings from recent literature, this chapter discusses the characteristics of current, conventional methods for ridership forecasts. Consequently, it aims at answering the following research questions:

- What are the characteristics of conventional methods for ridership forecasting and which problems are associated with it?
- How can the overoptimistic forecasts from conventional methods be explained?
- What are proposed solutions for the problem of overly optimistic forecasts?

Section 2.2 will show that past ridership forecasts have been, in general, overoptimistic. Subsequently, section 2.3 discusses the different causes of the overly optimistic results. Here, a distinction is made between technical causes for errors, optimism bias, and strategic misrepresentation. The case study in section 2.4 aims at comparing the findings from literature that are discussed in foregoing sections to the case of the circle-based method for ridership forecasts that is currently commonly used in The Netherlands. Proposed solutions for the inaccuracies, which should be the road to better forecasts, are discussed in section 2.5. A specific issue associated with the proposed solutions, namely the reference class problem, is presented and discussed in section 2.6. Finally, section 2.7 presents the conclusions along with a conceptual model.

2.2 Overly optimistic forecasts

Overestimated forecasts, which over-predict the success of new policy, are a well-known and widespread problem in transportation planning (Tal, 2008, p.3-4). In section 2.1 it has already been mentioned that this can have threatening financial consequences. The issue of large differences between forecast and actual performances of rail infrastructure projects, which are shown by large cost overruns and demand shortfalls, has been addressed in a number of recent studies (Mackett & Edwards, 1998; Flyvbjerg et al., 2005; Flyvbjerg, 2007; Flyvbjerg, 2008; Tal, 2008). The following quotations illustrate the consensus on the issue in and of itself and its severity as such:

> “[It] identified for large public procurement a demonstrated, systematic tendency for project appraisers to be overly optimistic” (Flyvbjerg, 2008, p.11).
>
> “There is clear evidence that forecasts of ridership are well in excess of what actually occurred” (Mackett & Edwards, 1998, pp.237-238).
>
> “Experience over the last three decades suggests that forecast of new policy aimed at changing travel behavior are usually overly optimistic” (Tal, 2008, p.4).
>
> “We conclude that traffic estimates used in decision making for rail infrastructure development are highly, systematically, and significantly misleading” (Flyvbjerg at al., 2005, p.144).

The best method for measuring the extent to which ridership expectations are met is comparing forecasts with actual ridership numbers (Tal, 2008, p.5). This kind of ex post analysis is, in fact the only way to measure the validity and accuracy of the forecast. There is simply no other way than to wait for actual data to become available and see whether or not the forecast had been correct or at least within certain error margins. A problem that arises with ex post analyses is, though, that it is in unpopular tool among politicians and certain planners due its judgmental nature (Spit & Zoete, 2005, p.87). As a result, ex post analyses on ridership forecasts are not always performed. In cases such are collected, the subsequent data do not always become publicly available. In other cases, forecasts can never be actually tested for their accuracy because the proposed station is never constructed at all, or the actual situation differs from the planned scenario used for the forecast. In other words, the quality of a forecast cannot be judged upon with disregard to the context in which it has been made (Tal, 2008, p.5; Toekomstonderzoek en Strategisch Omgevingsbeleid, 2000, p.88).

Despite the limitations mentioned above, research has been performed on the accuracy of ridership forecasts for rail projects. The referenced study is based on 44 urban rail projects from different parts of the world, including Europe and the United States. These have been compared to over 200 other transportation infrastructure projects. The sample size allows for statistical testing. Furthermore, due to the geographical distribution of the cases, comparisons between geographical areas were enabled. The group of 44 urban rail projects included both heavy and light rail and both underground and ground or elevated level (Flyvbjerg, 2007, pp.11-12). Based on the data from this database, remarkable conclusions have been drawn.

For one, rail projects suffer from the largest cost overruns with an average of almost 45 percent, compared to a 34 percent escalation for bridges and tunnels and 20 percent for roads. Three quarts of the rail projects under investigation had cost escalations of more than 24 percent. The average cost overruns in Europe equal 43.3 percent. Actual rail ridership is on average 51 percent lower than forecast. For a quart of the urban rail projects, actual generated forecast amounted to at least 68 percent lower than forecast. Only a tenth of the rail projects eventually achieved or went over the ridership number forecast (Flyvbjerg, 2007, pp.15-25; Flyvbjerg et al., 2005, p.133). The combination of high cost escalations and large demand shortfalls can be devastating, since they aid in the misrepresentation of rail projects to decision-makers, resulting in unviable projects being approved and executed. The combination of demand shortfalls and cost overruns is referred to as the double risk of urban rail (Flyvbjerg, 2007, p.18). Research by Flyvbjerg is in line with earlier, exploratory research from the United States that found that in most cases of forecasts used to justify urban rail projects, ridership was less than half of what was forecast. None of the projects that were examined had fully met, let alone topped, expected ridership numbers (Mackett & Edwards, 1998, p.238).

Despite the availability of better data, more powerful computers and consequent calculation capacity, and the possibilities for improving methodology for forecasting anytime during the past three decades, there has been no indication that either traffic forecasts in general, or rail ridership forecasts specifically, have become more accurate over time. In fact, statistical testing has brought forward no improvement in accuracy of traffic demand forecasts has taken place during the past 30 years and chronic overestimation is thus continuing until this day (Flyvbjerg et al., 2005, p.136; Flyvbjerg, 2006, p.6).

2.3 Causes for inaccuracy in ridership forecasts

The conclusions that rail forecasts suffer from chronic overestimation and no improvements have been made over the past three decades raise the questions how that it is possible that planners have not been able to achieve higher accuracies and what the causes for the imminent inaccuracies are. Based on consistencies found in literature, this section discriminates between technical causes for errors in forecasts, optimism bias, and thirdly strategic misrepresentation. Each is discussed in a designated subsection that examines the causes, the likelihood of occurrence, and the effects.

2.3.1 Technical explanations for forecast errors

Various sources identify the use of erroneous, obsolete, or irrelevant data and the application of inadequate methodology and models as expected causes for errors in ridership forecasting (Tal, 2008, p.3; Flyvbjerg, 2009, p.350; Mackett & Edwards, 1998, p.238). These causes can be grouped under technical explanations for inaccuracy. Other examples of technical explanations include honest human mistakes, lack of experience, and other inherent problems in forecasting such as a large degree of uncertainty about the future and random errors (Flyvbjerg, 2009, pp.349-351). It is even argued that the use of outdated and inappropriate data on the one hand and the use of improper models on the other are the two primary causes for inaccurate forecasts (Vanston & Vanston, 2004, pp.34-36).Tal emphasizes that forecasting human behavior, such as the choice between different modes of transportation, is by nature exposed to a high degree of uncertainty (Tal, 2008, pp.3-4).

According to this explanation, ridership forecast inaccuracies can be reduced with better data, more complex models, more experience among planners, and more checks for human errors. Nevertheless, as shown by Flyvbjerg (2006, 2007, 2008, 2009) and Flybjerg et al. (2005), 30 years of ridership forecasting has not lead to higher accuracies. A period of three decades however, should be long enough a period to have resulted in at least some improvement. Furthermore, human errors, incorrect data and random errors should have also resulted in underestimated forecasts. This, in turn, would have resulted in a normal distribution of errors. The fact that this is far from reality implies that there are other causes or reasons that have led to the notoriously overoptimistic forecasts for rail ridership.

2.3.2 Optimism bias

People's tendency to overestimate their own abilities is one of the primary sources for them to generally be highly optimistic for most of the time. Their positivism causes them to make decisions based on delusional optimism rather than based on rational weighting of pros and cons. As a result, costs are underestimated and gains are overestimated, as discussed in section 3.2. The optimism that forecasters have by nature affects their forecasts in a way that outcomes tend to be overly optimistic, because the possibilities of random events are ignored or underestimated. This is known as the planning fallacy (Lovallo & Kahneman, 2003, p.58-59). The effect that this cognitive predisposition has on forecasts is also referred to as optimism bias (Flyvbjerg, 2006, p.6). The existence op optimism bias in Dutch planning is acknowledged by Bakker and Zwaneveld (2009, p.59).

When forecasters or planners make forecasts for a new train station or another infrastructure project, they usually have an original rough-draft plan that has been drawn up by a local or regional government that believes constructing the project under investigation will aid local economy and improve accessibility. If the government had believed the contrary, that the project would be no use whatsoever, then they would not have come to the forecaster in the first place. This implies that the rough-draft plan, which now serves as a starting point for the forecasters, already has a positive judgment or at least a positive expectation put to it by the party that contracts the forecasters. This seemingly unproblematic procedure is objectionable however, because the use of the rough-draft plan will skew further outcomes toward additional optimism due to its initial optimism. This type of optimism bias, which is a prevalent type, is the outcome of so-called anchoring (Lovallo & Kahneman, 2003, p.60).

Optimism bias is a psychological explanation for cost overruns and demand shortfalls in rail infrastructure projects. Consequently, the existence of optimism bias would be able to account, at least partially, for the remarkable bias in ridership forecasts. These biases are believed to be omnipresent, but their ubiquity can be reduced with simple reality checks. When it is assumed that forecasters make overly optimistic forecasts by honest mistake, then the application of reality checks will help to reduce the errors in forecasts. It is expected that barriers to imply changes will be low, since planners who are prone to improvements will accept changes for the good (Flyvbjerg, 2008, pp.18-19; Flyvbjerg, 2009, pp.349-351). These barriers will be larger, however, when planners deliberately cause overoptimistic results.

2.3.3 Strategic misrepresentation

The intentional act of misinterpretation of data that leads to misrepresentation of results is known as strategic misrepresentation. It is the deliberate and systematic misstatement of fact to achieve organizational objectives in response to a clear set of incentives. Usually, strategic misrepresentation occurs in response to competition for financial resources (Jones & Euske, 1991, p.438). In case of rail infrastructure projects, planners and forecasters may strategically underestimate costs and overestimate ridership and benefits because this will increase the chances that their project is approved, funded and eventually constructed. Especially in the case of national funding that is based on a system of competition for the same budget between different projects, planners are aware that high costs reduce the likelihood a of project to receive national funding. Political pressure to secure funding from a regional, state, or national level is an additional reason for planners to purposely inflate benefits and hold back information about high costs (Flyvbjerg, 2009, pp.350-352).

The political and organizational pressures that result from the competition for project funding create an incentive structure that drives project managers and promoters to emphasize the gains and benefits of a project, while they de-emphasize the negative sides of the project, namely risks and costs. They rationalize this behavior because it is known that beneficial looking project is more likely to receive funding from higher-level governments. The latter should therefore stop offering categorical grants. These grants, awarded solely for the purpose of a project of a pre-defined type, create perverse incentives, as shown (Flyvbjerg, 2009, pp.358-359). Consider the situation in which multiple cities would like their city to be equipped with a new light rail system, for which subsidies from a national or supranational level are, although limited, available. The cities are thus in competition for the subsidies. In this example, the individual cities have an incentive for biased forecasts. In fact, if they do debias their forecasts, they will lose competition for funding to other cities unless they too debias their forecasts. In other words, accurate forecasts are often counterproductive whereas biased forecasts are the road to receiving funding. Project planners have confirmed that this is often the case and that strategic misrepresentation is used in these situations to secure funding for projects (Flyvbjerg, 2008, p.19; Flyvbjerg, 2009, pp.358-359).

From practice, there is specific evidence that strategic misrepresentation of facts has been used to get rail infrastructure projects approved. Particularly in the United States, the construction of certain light rail systems might not have been the outcome of rational decision-making on a local level, but rather the outcome of decisions based on untruthful forecasts. Cities such as Dallas, Minneapolis, and Sacramento seem to have chosen for high-tech, prestigious, glamorous light rail systems in cases where other solutions were, in fact, a lot more cost-effective. Although it now generates about 30,000 passengers on average weekday, a detailed study of the decision-making process of the light rail system in Sacramento, CA, abbreviated as LRT, brings forward that the manipulation of data and subsequent overstated benefits have led to choice of light rail over other options (Mackett & Edwards, 1998, p.2008). The Minneapolis, MN light rail connecting the city with Big Lake to the north, known as Northstar Commuter Rail, is a more recent example of a light rail project that is suspected to be the outcome of strategic misrepresentation (Levinson, 2011).

A colloquialism for strategic misrepresentation is lying. Strategic misrepresentation is, in fact, lying with data and numbers on costs, benefits, and risks. In the case of rail infrastructure, as shown by the examples of Sacramento and Minneapolis, lying can pay off or, at least, politicians, planners, and forecasters believe it does. According to the explanation of political pressure and funding, planners and forecasters purposely draw positive scenario's and camouflage or leave out numbers and details that uncover the downsides. This is, however, hard to prove. Lying, by definition, is based on making untrue statements to deceive others. Whether this has been someone's intention can only be established when that person's intention is known. It cannot be expected that planners and forecasters who initially have lied about a project, will later admit that they did so in order to ensure funding for a project, especially not if that means they run the risk of having to face legal actions. In cases of overly optimistic forecasts due to strategic misrepresentation, planners and forecasters are thus part of the problem and not part of the solution (Flyvbjerg, 2009, pp.350-358).

2.4 Case study: the circle-based method

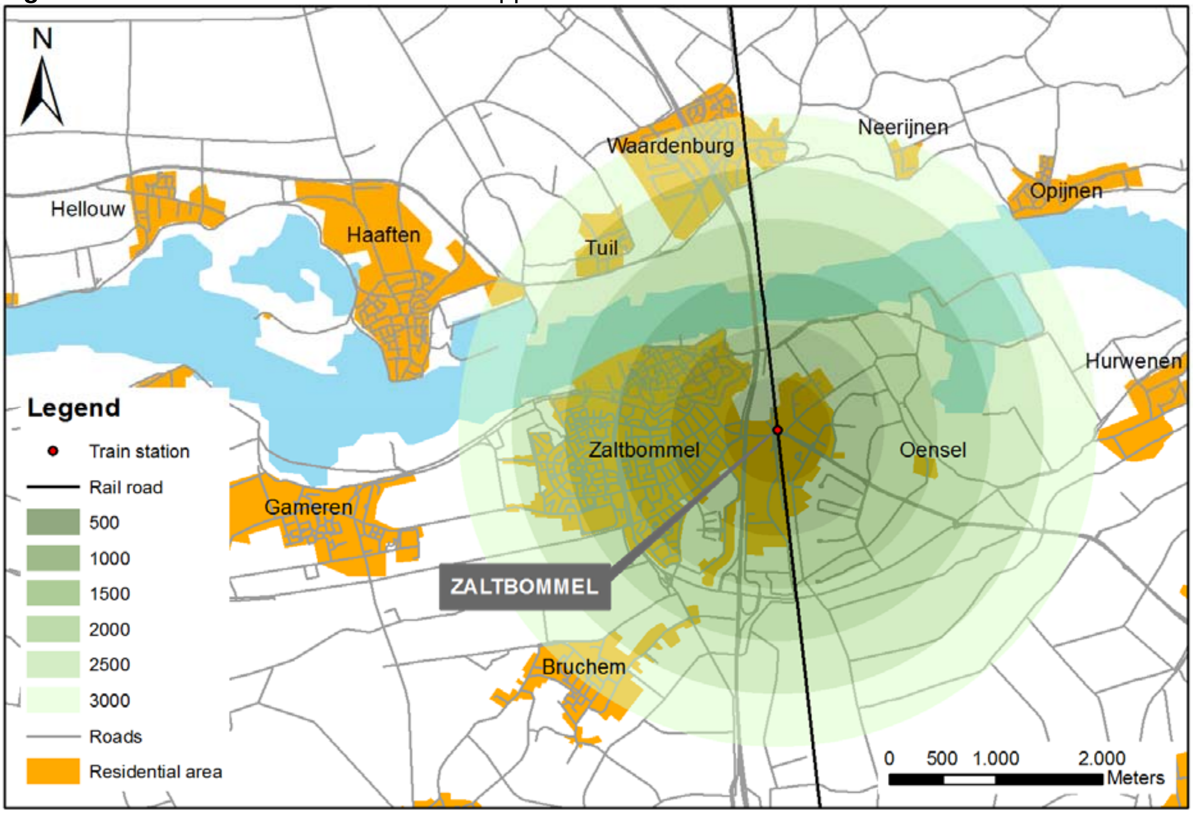
In sections 2.2 and 2.3, the problem of overestimation of ridership forecasts in conventional methods and its causes were identified and discussed. A method that is currently commonly used is the circle-based method. This method derives ridership forecasts form the number of residents and the number of jobs within crow-flight distance buffers with a 500 meter radius around the train station under investigation (see figure 2.1). Derivatives of this method are used by various traffic engineering and consulting businesses as well as passenger railway operators and ProRail in The Netherlands. The aim of this section is to perform a case study on the circle-based method to find to which extent the forecasts from the circle-based method are overly optimistic. Furthermore, the causes of overoptimism as identified in section 2.3 will be applied to this case to see to which extent they can be associated with the Dutch case.

2.4.1 Characteristics of the circle-based method

The circle-based method is based on the assumption that people who live closer to a train station are more likely to travel by train than people who live further away. Consequently, a relatively high percentage of the population within the first buffer is designated as train user. This percentage decreases with an increasing distance, which means that the furthest buffer is equipped with the lowest percentage. Usually, a total of five or six buffers are applied, which adds up to a 2.5 or 3 kilometer radius respectively. The same principle is used to derive the amount of train movements that are associated with local employment. It is assumed that a set percentage of employers within the circles will travel to their work by train when possible. In some examples, the circle-based method has also been applied to college enrollment. The circles are drawn in GIS, which is subsequently used to calculate the number of residents, jobs, and student enrollment within these circles. Final calculations, which result in ridership forecasts, are made in a spreadsheet.

Various traffic engineering and consulting businesses in The Netherlands have applied the circle-based method for ridership forecasting. Incidental examples include Inno-V for the stations Driehuis, Santpoort Nooord and Santpoort Zuid (Nanninga & Woolthuis, 2004), Arcadis for stations along the track between the cities of Maastricht in The Netherlands and Aachen in Germany (Arcadis, 2009), and Witteveen + Bos for the proposed train stations Baexum and Haelen (Witteveen + Bos, 2011). Goudappel Coffeng appears to be a more frequent user of the circle based method. Studies by Goudappel Coffeng that involve the circle-based method are, among others, proposed train stations in the province of Overijssel (Van Leusden & Wegewijs, 2009) and proposed station Zantpoort (Andriesse & Ebben, 2010). The circle-based method used by ProRail and NS is known as PINO. Though the basics of the model do not differ much from those used by the commercial businesses, PINO is a little more sophisticated since it also takes into account others modes of public transit that can be feedering or competing. For the purpose of the study, the focus of this case study will be on the forecasts made by Goudappel Coffeng's circle-based method and ProRail's and NS's PINO.

Figure 2.1: Circle-based method buffers applied to the case of Zaltbommel train station



Source: Goudappel Coffeng, 2011; Open Street Map, 2011; own adaptation, 2011.

2.4.2 Overoptimism of circle-based method forecasts

In section 2.2 it has been discussed that Flyvbjerg (2007) has found that, on average, rail ridership is 51 percent lower than forecast. To see whether such overly optimistic forecasts are also made for train stations in The Netherlands, a comparison is made between the outcomes of PINO and the actual ridership numbers generated after the construction of the stations. The results of the comparison are presented in table 2.1. Results of Goudappel Coffeng forecasts are not included because, with the exception of Maarheeze station, the stations studied have not been constructed. This makes it impossible to value the forecasts.

Table 2.1 shows that only two stations have exceeded the expectations in terms of average daily boardings. These stations are Tilburg Reeshof and Tiel Passewaay, with ridership overruns of 14.9 percent and 11.8 percent respectively. It is fair to state that the expectations for Almere Oostvaarders have also been met, considering a shortage of only 1.7 percent. The 16 other stations under investigation have generated ridership numbers that are lower than forecast. The smallest negative error is 17.7 percent for station Voorst-Empe, while the most overly optimistic forecast has turned out be for Utrecht Terwijde, where ridership is 82.7 percent lower than forecast. Actual ridership numbers for the 16 stations that have not met the expectations are on average 49.3 percent lower than forecast. When all 19 cases are considered, the average error is still -40.2 percent. When an error margin of ±25 percent is considered acceptable, there are still 14 forecasts that fail to meet the requirement. Ridership numbers for these 14 stations are on average 53.3 percent below forecast.

Table 2.1: Comparison of ridership forecasts by ProRail's PINO with actual ridership data (2008, 2009)

Railway station	Actual # of daily boardings	Forecast ProRail's PINO	Error of forecast (difference in %)
Tilburg Reeshof	1838	1600	+14.9%
Almere Oostvaarders	3439	3500	-1.7%
Den Haag Ypenburg	1238	2150	-42.4%
Arnhem Zuid	1247	3900	-68.0%
Helmond Brandevoort	642	2050	-68.7%
Utrecht Terwijde	1384	8000	-82.7%
Amersfoort Vathorst	1633	2500	-34.7%
Tiel Passewaay	1230	1100	+11.8%
Utrecht Zuilen	1397	2000	-30.2%
Krommenie-Assendelft	4169	6700	-37.8%
Apeldoorn De Maten	636	1750	-63.7%
Apeldoorn Osseveld	773	1500	-48.5%
Gaanderen	339	650	-47.8%
Twello	1330	1750	-24.0%
Voorst-Empe	288	350	-17.7%
Groningen Europapark	862	3500	-75.4%
Purmerend Weidevenne	1578	2125	-25.7%
Eygelshoven Markt	149	400	-62.8%
Heerlen De Kissel	419	1000	-58.1%

Source: Projectteam Toepassing Norm ProRail, 2009

2.4.3 Causes of overly optimistic results

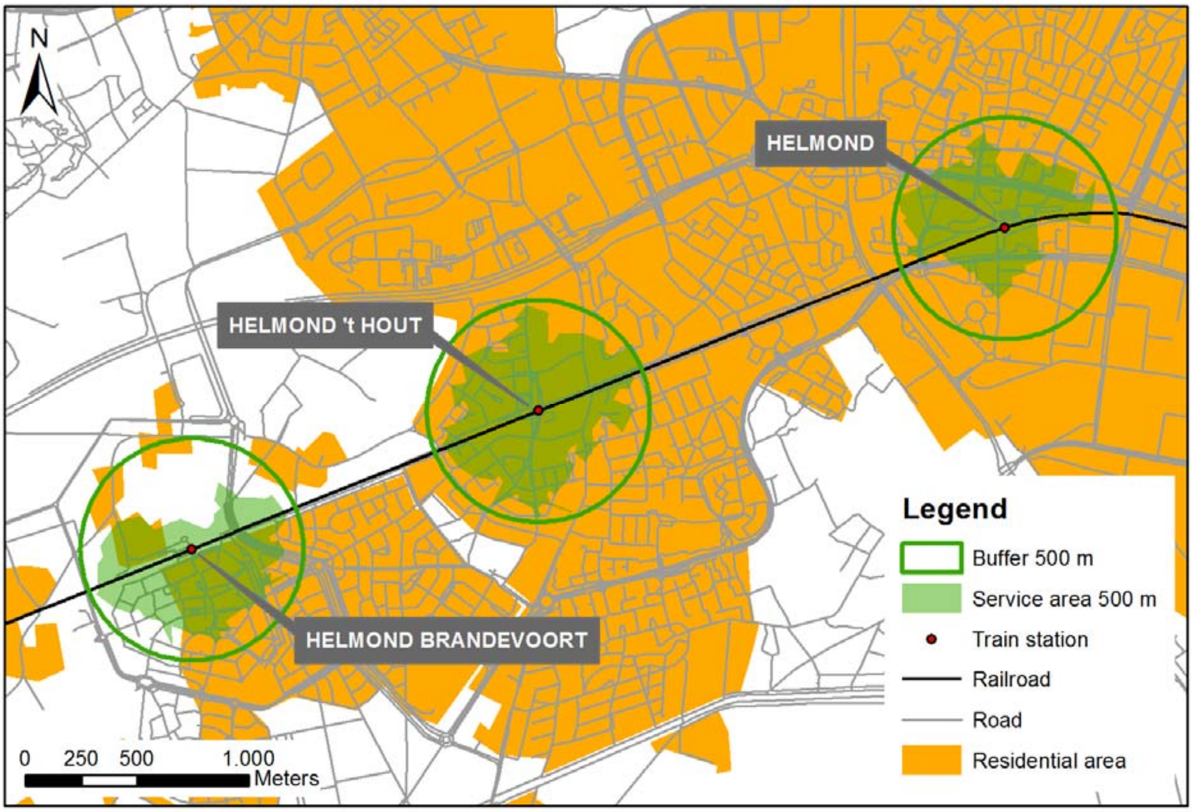
Section 2.3 has discussed the causes of overly optimistic ridership forecasts. It discriminates between technical causes, optimism bias and strategic misrepresentation. All three are associated with overoptimism in ridership forecasts, but the question is to which extent each cause affects the results. This subsection aims at answering this question for the forecasts made by PINO.

Newly opened train stations usually need a few years to grow to their full potential. Precise data are not available, but it is expected that new stations generate growth in daily boardings for four through seven years before their ridership numbers stabilize. The forecasts presented in table 2.1 have taken this start-up issue into account, considering that various forecasts have been calibrated for 2010 or 2015 (Projectteam Toepassing Norm ProRail, 2009). As a consequence, not all cases of comparison between the forecasts and the actual data can be used to draw final conclusions. Another limitation is the fact that the forecasts are often based on assumptions on future developments. If contemporary spatial plans include the construction of new housing units for example, this will be included in the forecast. When the construction of additional units is subsequently canceled, it should be noted that the forecasts become less applicable or even totally inapplicable, since they were made for a different actual situation. Utrecht Terwijde station provides an example of a forecast that was based on a situation that differs much from the actual situation. At the time of the forecast of 8000 daily boardings in 2011, it was assumed that 18,000 dwellings would have been constructed around 2009. The apparently overly optimistic forecast can partially be due to the absence of the expected construction of new dwellings. This can be defined as either human misjudgment of the spatial plans, as inaccuracies in the data used, or as an inherent problem of uncertainty in forecasting. Regardless of the choice between these three, it can be approached as a technical cause of error.

Of the technical causes of errors in forecasting, data and model limitations and inaccuracies are most frequently mentioned (see subsection 2.3.1). It may be presumed that, other than data inaccuracies due to uncertainties about the future, population data used in the circle-based method is usually correct. These data can be acquired through the Dutch national bureau of statistics (CBS) and is therefore regarded accurate. This should furthermore ensure consistency in definition, acquisition and geography. It does not guarantee, however, that these data are applied correctly by forecasters, nor that no errors are made in calculations within a spreadsheet or GIS environment. Employment data are not as easily available as population data through CBS. Local and regional governments might be better sources for employment data, but the exact sources of such data used in circle-based method calculations remains unclear. A closer look at calculations by Goudappel Coffeng shows that employment usually only has minor effects on the total expected number of daily boardings. Taking these findings into consideration, one can conclude that data quality, or rather a lack of quality, is not a major source of the overoptimism in ridership forecasts.

The circle-based method itself could be a cause of the overly optimistic results. The two aspects of the method that can be directly associated with the outcomes are the parameters and the way in which the stations' coverage areas are calculated. The former are usually set based on a nearby existing train station that is regarded by the forecaster as similar. Calculations by Goudappel Coffeng for Zevenaar Oost, for example, used parameters derived from the existing station of Zevenaar, based on the assumption that both stations will attract the same kind of people since they are located within the same municipality. A similar approach has been taken in the calculations for Nijkerk Corlaer, in which the existing station Nijkerk served as a reference. This approach can be criticized for various reasons. First, the choice of the reference station is arbitrary and hardly answered for. It is not said that the conditions is a suburban area are similar to those in the downtown area of the same city. Especially in the case of a newly constructed residential area, the makeup of the population may be totally different from that in the rest of the city. Secondly, the use of a single reference makes the model susceptible to statistical outliers and thus improbable results. Furthermore, the choice of a reference station nearby does not guarantee similarity. There are many more factors that will affect ridership other that geographical proximity (see chapter 3). They way these flaws might relate to the overoptimism in the forecasts is through the aforementioned start-up time a station needs. The existing stations used as references have usually been there for at least several decades, which means that close residents have long grown accustomed to the train as a mode of transportation. This is obviously not the case in newly developed suburban areas that are equipped with a new station. Therefore, less people will use the train, which means the expected volumes are not realized for at least five to ten years.

Figure 2.2: Stations' catchment areas as 500 meter radius buffers versus network-based service areas in the case of Helmond



Source: Goudappel Coffeng, 2011; Open Street Map, 2011; own adaptation, 2011.

The second disputable aspect of the circle-based method is the determination of the stations' service areas. As discussed earlier, these are currently based on Euclidean distance. Euclidean, or crow-flight distance is well applicable to helicopter accessibility analyses, noise pollution, and proximity to wells in agriculture, but is unfit for urban, human applications. One can easily imagine that the Euclidean distance to a train station is hardly ever equal to the actual distance a person has to travel to get to the train station. In fact, the real distance along the road network is always longer due to everyday obstacles in the cityscape such as buildings, waterways, impermeable roads and private property. GIS network analysis capabilities provide the means to calculate service areas along the road network. This provides more accurate, veracious coverage areas than Euclidean distance (see figure 2.2). Where the former is associated with overestimation, the latter has been shown to lead to systematically better forecasts of transit ridership (Gutierrez & Garcia-Palomares, 2008, pp.480-502).

It is not likely that the technical causes fully account for the overly optimistic ridership forecasts. Various attempts have been made to adjust the current models in order to gain more accurate forecasts. It is justified to look for other causes since these attempts have neither led to more accurate results nor to a more even distribution of errors. Optimism bias and strategic misrepresentation may therefore also account for parts of the overly optimistic forecasts. Optimism bias, for starters, is likely to take place in ridership forecasting in The Netherlands. The expectation that a certain station will have a positive effect on a location's accessibility and, consequently, benefits local economy is usually a starting point for further analysis. Local policy makers are thus already infected by certain expectations regarding the performance of "their" station when they approach forecasters and planners. The positive starting point may thus cause forecasters to, subconsciously, skew the results of subsequent analyses towards overoptimism. Optimism bias can be apparent to both ProRail and commercial businesses that perform forecasters as contractors for governmental organizations.

As discussed in subsection 2.3.3, strategic misrepresentation usually occurs in response to competition for resources. The national government of The Netherlands provides grants for the construction of new train stations as part of a special program for this purpose (Programma Aanleg nieuwe stations). In 2009, the total budget for this program was 77 million euro. A single station is subsidized up to a maximum of 6.3 million euro (Ministerie van Verkeer en Waterstaat, 2009, p.106). This means that there is a budget for grants for the construction of twelve new train stations, assuming that all will be subsidized maximally. When the 90 or so proposed train stations are considered, it can be noticed that competition for financial resources is the case for new train stations in The Netherlands. Because new train stations have to meet certain criteria regarding economic viability in order to be eligible for the national grants, there is an incentive for local decision-makers to depict their projects as such. It is therefore conceivable that strategic misrepresentation is occurring in The Netherlands. This does not, however, account fully for the overly positive forecasts by ProRail, since it has no such incentives for individual stations. It should therefore be concluded that the overly optimistic ridership forecasts for train stations in The Netherlands are the result of the co-existence of model limitations, optimism bias, and strategic misrepresentation.

2.5 Proposed solution: Reference class forecasting

The American Planning Association (APA) has acknowledged the persistent problem of inaccurate, biased forecasts in public projects, including rail infrastructure projects. APA has encouraged all planners to embrace efforts to achieve more accurate forecasts for public projects and to take an active approach in doing so. As a solution for overly positive forecasts and cost underestimations, APA has recommended the use of a method called reference class forecasting (APA, 2005). This method, RCF, is an objective method that avoids psychological, organizational and financial sources that are associated with optimism in forecasts (Lovallo & Kahneman, 2003, p.61). Thus, RCF is thought to be a sufficient method for bypassing the problems that arise from optimism bias and strategic misrepresentation. The method of RCF is also referred to as the outside view.

In RCF, or the outside view, forecasters should completely ignore the details of the case at hand, such as its objectives, the resources available to it, expected scenarios, and possible hiccups. Furthermore, forecasters need to refrain from creating images of the expected future developments in their minds. Instead, RCF is based on historic precedent brought forward from a group of similar cases, which together make up a reference class. From, the reference class, expectations for the investigated case can be derived based on statistics. In order to enable extrapolation from the reference class, it needs to include a minimum number of 20 cases to draw statistically acceptable conclusions. Simultaneously, the reference class should be narrowed down to statistically similar cases to make it comparable with the subject at hand (Lovallo & Kahneman, 2003, p.61; Flyvbjerg et al., 2005, pp.140-141; Flyvbjerg, 2006, p.8).

Contrary to the outside view is the inside view. The principles of the inside view are characteristic for conventional methods of forecasting, including those used in rail infrastructure projects. By taking an inside view, forecasters focus tightly on the specific or unusual features of the project and try to use these to predict their influence. Subsequently, predictions for the project, such as expected ridership numbers, are made based on these findings. Taking the inside view is the conventional way of forecasting. Therefore, planners' and forecasters' preference for this approach is understandable. Yet, since research has shown that the use of the outside view yields better results than the use of the inside view, the use of the outside view is recommended (Flybjerg et al., 2005, p.141; Flyvbjerg, 2006, p.8). The difference between the inside view and outside view can be illustrated with an example. Consider a group of university freshmen who are asked about their expected performances in the first year of their academic endeavors. On average, they expected to be performing better than 84 percent of their peer students. Realistically, this should have been 50 percent, since that is the actual average. A control group was asked about their entrance scores prior to the question about their academic performances. This sequence of questions reduced the average expectation to perform better than peers to 67 percent. This is still overly optimistic, but the percentage is considerably lower. The former group applied an inside view, whereas the latter group was forced into considering findings from a reference, which resulted in more realistic results (Lovallo & Kahneman, 2003, p. 61; Flyvbjerg, 2006, p.8). Though this is a highly simplified example of the use of the outside view, it does provide an indication of the higher accuracy that can be achieved.

2.6 The reference class problem

The process of RCF consists of taking three steps. The first step is to identify a reference class that contains past projects which are relevant to the case at hand. The class needs to contain a minimum number of cases to allow statistically meaningful analyses but it should simultaneously be narrow enough to ensure that all cases share sufficient common features to enable comparison. Secondly, the distribution for the selected reference class needs to be assessed. The third step consists of comparing the project at hand with the distribution of the reference class to establish a forecast (Flyvbjerg et al., 2005, p.140). The first of the three steps can cause problems. The first problem arises when there is alack of precedents (Lovallo & Kahneman, 2003, p.63). A major problem, which is eminent in RCF, has to do with the choice of the reference class. Though this problem is not digressed on in literature on RCF, it can still be identified from the examples used. Regarding the choice of a relevant reference class, Flyvbjerg et al. (2005) provide an example of the construction of a new subway line where a reference class of comparable projects should be established. "This could be the urban rail projects included in the sample for this article" (Flyvbjerg et al., 2005, p.141). It could, in fact, be that specific set of rail projects. But it could just as well be a set of subway projects that excludes light rail and heavy rail projects. Or it could be narrowed down to subway projects on the continent of the project at hand. Other options for reference classes could be based on other characteristics of the subway project, such as length of the line, number of terminals along the line, or the frequency of service. In actuality, the project can be incorporated into many different reference classes, for which different probabilities exist. This issue is referred to as the reference class problem (Hájek, 2006, pp.2-3).

The well-known problem of the reference class is often elucidated by the Shonubi example. Charles Shonubi is a Nigerian citizen and a legal resident of the United States, where he was arrested at JFK international Airport in 1991 for smuggling heroin. There was very little doubt about his guilt, since he carried over 100 balloons containing heroin in his digestive system. Furthermore, there were indications and considerable evidence that Shonubi had been carrying heroin across the border on several previous occasions. Consequently, Shonubi was convicted and sentenced to twelve years and seven months of imprisonment. The sentencing was based on the assumption that he had carried similar amounts of heroin on seven previous occasions. This was thought to have accumulated to more than 3,400 grams. After an appeal, the trial was sent back to the district court because the reasoning was not apt and the evidence subsequently insufficient. In response, data from a reference class, which consisted of data on 117 other Nigerian drug smugglers arrested at JFK International Airport during the period between Shonubi's assumed first attempt and his arrest. It was concluded that he had smuggled at least 2,000 grams of heroin on his prior attempts with 0.99 certainty. The evidence was presented as uncontroversial while, in fact, it was not. Shonubi is a member of an infinite of reference classes. The prosecution could have used statistics on drug smugglers at JFK regardless of citizenship, Nigerian drug smugglers at all US airports, or all drug smugglers in general. They could even have used intuitively unapt references to all Nigerian airline passengers or toll booth collectors, which was his day-to-day profession (Cheng, 2009, pp.1-2; Colyvan et al., 2001, pp.168-171).

The question that should have asked during Shonubi's trial was whether the chosen reference class was the appropriate one. Choices for other reference classes could heave been equally reasonable. Yet, each reference class will have its own probability and will therefore lead to different outcomes (Hájek, 2003, p.3). Out of all possible reference classes, the most suitable option should be chosen. Unfortunately, the optimal way in which to achieve the best comparison group is yet to be established. Since the aim of this research is not to provide a solution to the reference class problem, this issue will not further be diverted into. Instead, a solution or, at the very least, a workaround for the reference class problem in the case of ridership forecasting needs to be found. Findings from a literature research on factors influencing ridership, which are presented and discussed in the following chapter, will yield a number of variables that are associated with train ridership. One, or a combination of a small number of these will lead to the best possible division of all stations into distinct, internally homogeneously classes. One possible approach for dividing all cases into reference classes is to intuitively define the most relevant factor and subsequently divide the cases based on their values for the respective factor. This is a suboptimal approach, since it is subjective and thus arbitrary. Another option is that of trial-and-error. This method is also suboptimal, since it involves a time-consuming repetition of tests for each individual factor and, possibly, each combination of identified factors. In this research, a combination will be used to define the reference classes best applicable to the Dutch case. Intuitively, based on statistical testing, and based on indications provided by literature, a number of factors will be phased out as diversifying for the reference classes. The factors that will remain are subsequently tested for significance on ridership and their suitability to be used as a basis for the creation of reference classes of train stations in The Netherlands.

2.7 Conclusions

In this chapter, the characteristics of ridership forecasting methods have been discussed. An important finding is that that all seem to suffer from chronic overestimation of ridership. Literature strongly indicated that overly optimistic forecasts are to be expected in The Netherlands. This has been confirmed by a brief analysis of the forecasts for 19 recently opened train stations. The expectations were, on average, far from the eventually generated numbers. Three different types of causes for the overly optimistic results can be identified. Firstly, technical causes that are associated with overoptimism include the uncertainty about the future that is inherent in modeling, human errors, and model limitations. The latter is thought to relate to the use of Euclidean distance rather than road network buffers. Secondly, optimism bias, which is the natural human tendency towards optimism, is thought to also affect the forecasts. Finally, strategic misrepresentation, or simply lying, is identified as an important reason for overly optimistic forecasts. Competition for financial resources is thought to be a driver for strategic misrepresentation, since it provides the incentives to depict a project more positively than it actually is.

Reference class forecasting (RCF), or the outside view, has been proposed as a solution to the overly optimistic forecasts generated by conventional methods. Instead of focusing on the details of the project at hand, RCF makes use of a class of similar cases from which a forecast is derived based on the common characteristics. This method is proven to yield more accurate forecasts than conventional methods. To enable statistically acceptable forecasts, the reference classes need to be composed of a significant number of cases. At the same time, the cases need to be alike to enable comparison. A problem that is associated with RCF that cannot be neglected is known as the reference class problem. There is no straightforward answer to the question of which variable should be used to base the reference classes on. Theoretically, every case has an infinite number of reference classes it belongs to.

3 Theory on ridership generation

3.1 Introduction

Whether or not and, in case they do, to which extent people make use of the train as a mode of transport depends on various conditions. It is known that 38 percent of the Dutch population never makes use of the train (NS, 2011). In contrast to those who never travel by train are those who are bound to public transportation for their journeys. These are known as the so-called transit dependent or captives, who are adults without a driver's license or those without access to a private car. Transit dependents make significantly more use of public transportation than others (Bakker & Zwaneveld, 2009, pp.37-39). This indicates that car ownership has a negative effect on train ridership. Among other aspects that are expected to have an effect on train ridership are household type and composition, income, level of education, and degree of urbanization of the residence. The aim of this chapter is to find indications or evidence from recent literature on the relevance of variables which are expected to have a significant effect on ridership generation. In order to do so, this chapter discriminates between user characteristics, train station characteristics, and characteristics of the stations' built environments. This chapter consequently seeks to answer the following research questions:

- What are, generally speaking, the features of train passengers and how can the transit dependent traveler be described?
- Which factors are determinants for passenger railway transportation, apart from passenger characteristics, and consequently play a part in ridership forecasting?
- What are determinants for ridership from the built environment of train stations and how do they add to the number of daily boardings?

This thesis focuses on the development of a method for ridership forecasting for train stations in The Netherlands. Because the Dutch situation differs significantly from those in other countries, findings from international literature should be applied with care to this case. For example, the majority of the access journeys to train stations in The Netherlands are made by bicycle, while statistics from Sydney, Australia indicate that the majority of train users walk to the station (Transport & Population Data Centre, 2003, p.3). Another example is provided by the grid street patterns which are characteristic for various North American and Asian cities. Gird street patterns are thought to have a positive effect on train ridership, but one should bare in mind that these are far from prevalent in The Netherlands (Akayima & Okushima, 2009; Cervero & Kockelman, 1997).

3.2 Train users' characteristics and behavior

Proximity to a train station is a major determinant for train usage. It is generally acknowledged that people who live closer to a train station are more likely to travel by train (Keijer & Rietveld, 2000; Akiyama & Okushima, 2009). In The Netherlands, the average distance to the nearest train station for residents is 4.5 kilometer, while the mode is only about 1.3 kilometer. The percentage of the Dutch population living outside a 10 kilometer parameter is 8.4 (Keijer & Rietveld, 2000, p.232). Research has shown that people's propensity to travel by train decreases heavily when their access journey to the station is over 3.5 kilometer (Keijer & Rietveld, 2000, p.233). Even within the 3.5 kilometer threshold it is hard to generate additional ridership, since it is extremely difficult to get people to abandon their cars for public transportation, which is illustrated by the analogy of the impossibility of forcing Mickey and Minnie Mouse to divorce (New York Times, 2011).

The proportion of people who do use public transportation, the train in particular, cannot simply be captured with generalizing statements about them. In fact, train users can be subdivided into different classes based on their needs or their styles of mobility. Hagen et al. (2005) discriminate between seven groups of passengers by the demands and needs they have with regards to traveling by train. A distinction is made between convenience seekers, life enrichers, status seekers, retreaters, functional planners, capacity seekers, and socializers (Hagen et al., 2005, pp.11 & 20). Another distinction is based on the reasons and purposes of people's trips by public transportation. These groups are named caring workers, traditional commuter, high school and college students, family and friends visitor, relaxer, care-taker, relishing worker, and the public transportation professional (Verburg et al., 2005, pp.8-9). Both ways of discriminating between train users seek to identify target groups based on their characteristics to enable specific product development and adjustments. The distinction by Verburg et al. is the most useful is this case, because it is based on general personal characteristics that can be measured, rather than specific personal preferences and needs.

Characteristics that are associated with train ridership are household type, household size, household composition and the presence of young children, work and study, level of education, income, degree of urbanization of the place of residence, age, and car ownership (Verburg et al., 2005, p.6). It is shown that the use of public transportation is particularly popular among households with a strong focus on professional work. The train is more popular among these households than other types of public transportation. Nevertheless, public transportation user groups show an overrepresentation of singles, elderly, and to a lesser extent, young adults (18 through 30 years) and members of households with a less than average income. The group of young adults is mainly composed by students who are in the possession of public transportation student passes (Verburg et al., 2005, pp.11-30). A quart of all kilometers traveled by train is made by students who own a public transportation student pass. Those with a lower than average income make relatively more use of public transportation than those with higher incomes, but the former still travel two thirds of their total kilometers by car (Bakker & Zwaneveld, 2009, pp.44-45). Though those with lower incomes make relatively more use of public transportation, those with higher incomes make more use of public transportation in absolute terms (Bakker & Zwaneveld, 2009, p.45).

3.2.1 Comparison with car users

The fact that those with higher incomes make up the largest group of users of public transportation in absolute terms does not mean that they travel less by car. To the contrary, those with higher incomes also have an absolute high stake in kilometers traveled by car, period. This example indicates that it is not always easy and can be confusing to compare different findings. Nevertheless, an attempt can be made to characterize the average train user by comparing him to the averages of other mobility modes, particularly the car. It is known, for starters, that public transportation users on average undertake the longest enduring daily trips and also have the longest lasting activities in between their access and egress journeys (Verburg et al., 2005, pp.6-7). Furthermore, research has shown that public transportation users spend more time on work than do users of the car. In addition, public transportation users spend a relatively large amount of time on education, which is to a large extent explained by the large amount of public transportation student pass holders on trains and other modes of public transportation. The public transportation user is further characterized by his destination, which is more urbanized than those of car users. Public transportation users have fewer young children than car users, are generally more highly educated, and their places of residence are more urbanized than those of train users (Verburg et al., 2005, pp.7&29).

3.2.2 Transit dependent travelers

An important distinction can be made among public transportation users. On the one hand, there is a group of travelers who have the possibility of choice between different modes of transportation for their trip. Usually these are both a car and public transportation. On the other hand, there are those who cannot make a choice between a car and public transportation. This group is known as transit captives or transit dependent travelers. The distinction between both groups has somewhat of a fuzzy boundary, since a person can be a traveler of choice on one day while he is transit dependent on the next, for example when a two person household shares a single car (De Beynen de Hoog, 2004, p.5). Contrary to the transit dependent travelers is the group of car dependent travelers. This group is characterized by their inaccessibility to public transportation. This can be the case in remote areas, for those who leave or arrive at times that are not serviced by public transit, or for those who face financial or physical limitations (De Beynen de Hoog, 2004, p.9). Car captives are overrepresented in the lower income classes.

A problem arises when trying to define the transit dependent traveler. It is said to be a person whose only option for transportation is public transportation. Here, it is the question what that person personally sees as an option. Something that can be an acceptable option for person A, can be absolutely out of the question for person B. This can be illustrated by the example where person A is willing to walk 15 minutes to the nearest bus stop while person B would never do that, for whichever reason. Defining the transit dependent traveler is, in this perspective, a subjective rather than an objective matter (De Beynen de Hoog, 2004, p.5).

To overcome the matter of subjectivity, the group of adults who are not in the possession of a driver's license can be regarded as an objective substitute. This is not a perfect substitute, because one can still be driven by someone else and because one can own a driver's license without owning a car, but it can serve as a viable approximation. Not having a driver's license is overrepresented among those between the ages of 18 and 25 and those over the age of 67. The total group of people lacking a driver's license is composed of 25 percent young adults (18-25, evenly divided over both sexes) and for one thirds of elderly (67+, 80 percent female). Among people between the ages of 25 and 67, not having a driver's license is less prevalent than average. Still, about 40 percent (of which 70 percent is female) of all people without a license is from this age group. An analysis of this group based on social position reveals that housewives, the unemployed, and those who receive incapacity benefits are overrepresented (Bakker & Zwaneveld, 2009, p.38).

3.2.3 Changing demographics in an aging society

One of the goals of public transportation that is often mentioned is to help senior citizens with their mobility needs. In an aging society, such as The Netherlands, it is expected that the growth of the group of elderly (over the age of 65) will have only minor effects on train ridership since retirees lack the school and work motives to travel (Bakker & Zwaneveld, 2009, pp.40-41). Until now, people over the age of 65 spend less time on public transportation and travel fewer kilometers by public transportation compared to other age groups. This is not expected to change, but specific consequences may include a rise in public transportation trips for social of leisure activities and a decrease in work-related travel during peak hours (Verburg et al., 2005, pp.22-23). This is in accordance with findings from research on rail ridership in the city of Seoul, South Korea (Akiyama & Okushima, 2009, p.11).

Other demographic changes expected for the near future, besides an aging society, are increases in the amount of households with higher than average incomes and the amount of single person households. The extent to which these changes will affect ridership is not fully understood (Verburg et al., 2005, p.20; Bakker & Zwaneveld, 2009). Other trends discussed in literature include an increase in car ownership, and consequently an increase in car use and a decrease in pedestrian traffic and public transportation use in the city of Seoul. The decrease in public transportation use is relatively small in bus, but significantly large in rail ridership (Akiyama & Okushima, 2009, p.8). These trends from South Korea are, however, not perceived in The Netherlands. Public transportation already plays a relatively minor role in total traffic in The Netherlands (Verburg et al., 2005, p.29), while it is a major means of transportation in the city of Seoul.

3.3 Train station and network characteristics

Train stations can function as generators for economic growth. Specific recent examples include Amsterdam Zuid/WTC and Rotterdam Alexander, which have shown to be attractive sites for the development of offices and other businesses. At the same time, other commuter train stations that have been opened in recent years are lagging in ridership compared to the expectations and do not or barely generate new developments (Vergouwen & Baggen, 2004, p.4). The differences in daily boardings and spatial developments between train stations which are potentially similar, brings up the question of how the different outcomes can be explained. The aim of this section is consequently to find variables that are associated with train ridership and are expected to affect it.

3.3.1 The effect of an intercity status

The level of attraction that a train station has on potential users depends on various factors. First, and perhaps foremost, people are interested in the destinations that can be traveled to and the time it takes to get there. Large intercity stations attract larger crowds than suburban stations because intercity stations provide access to the main rail network and consequently the largest cities of the country. Furthermore, travel times are usually lower by direct intercity train than by regional train and frequency of trains is much higher at the former. A high frequency reduces average waiting time, which improves people's perception of the trip as a whole. The impact of an intercity train station status and the consequent improved frequency of trains has been shown by the results of the train station in the town of Best. This station generated 550 additional boardings per day after it received the status of intercity train station. Two additional intercity trains in each direction every hour and additional parking spaces (park-and-ride) were part of the station's upgrade (De Vos, 2009, p.12). Due to the upgrade to intercity station, other train stations became accessible within shorter time. The trip between Eindhoven and Schiphol took four minutes longer as a consequence of the additional stop at Best, but nevertheless ridership per train increased on this route (De Vos, 2009, p.14).

3.3.2 Modal access

The case of Best included extra parking spaces. This may have had a positive effect on ridership. It is suggested that the availability of park-and-ride facilities has a strong positive association with rail ridership, since commuters with a residence outside of walking distance from a train station reach it by car. Research shows that every additional park-and-ride space yields 0.77 boardings in the United States (Kuby et al., 2004, pp.234&242). It can be argued whether park-and-ride nets the same results in The Netherlands. It is known that the bicycle is an important access mode in The Netherlands (Givoni & Rietveld, 2007; Keijer & Rietveld, 2000), which could mean that the effect of car parking facilities is much smaller than in the US. Instead, improving accessibility for bicycles could draw more passengers into the train, since about one in three already uses the bicycle as a mode of transportation for the access journey (Keijer & Rietveld, 2000, p.234). Furthermore, it can be argued that free parking has a negative effect on ridership, since it can add to the attraction of the automobile. Parking costs could therefore have a positive relationship with ridership, since higher costs for parking in train station areas can discourage the use of the car and simultaneously draw people into public transportation (Kuby et al., 2004, p.227).

Other modes of public transportation, being bus subway and light-rail, cumulatively account for 26.7 percent of the access journeys to and 34.6 percent of the egress journeys from train stations in The Netherlands (Givoni & Rietveld, 2007, p.359). Good connections with other types of public transportation are named as one of the characteristics of a successful train station (Van der Bijl & Maartens, 2010, pp.12-15). Research has found that the frequency of buses is major factor in ridership generation (Currie et al., 2011, p.548). This suggests that co-location of a number of stops and stations of other types of public transportation at a train station and high frequencies of stopping modes will add to the number of daily boardings. It should be mentioned, however, that perpendicular or parallel bus lines might have a negative effect on ridership, since they might draw passengers away from the train and into the bus.

3.3.3 Railroad network characteristics

Subsection 3.3.2 has discussed the relevance of a high frequency of feedering public transportation for daily boardings of train stations. A higher number of buses, street cars and subways is a driver for more ridership for a train station. Similarly, it is thought that a higher frequency of trains at the station itself also generates higher ridership numbers. In fact, high service levels are identified as one of the main drivers of transit ridership (Currie et al., 2011, p.546). The effect of a higher frequency of trains has also been suggested by the case of the station of Best (see subsection 3.3.1). The fact that there is a positive relationship between service levels and the number of daily boardings is not surprising, since a higher frequency reduces the average waiting time and possibly the transfer times. Since waiting is valued 2.5 times worse and transfer time 1.75 times worse than time spent within the transit mode (Hof & Koopmans, 2006, p.8), improving service levels are hypothesized to have a positive relationship with ridership in The Netherlands.

When a railroad line is home to a relatively high number of train stations, there will be more potential passengers living within walking and cycling distance of stations along that line. It also enables passengers to visit a high number of places. Hence, it should have a positive effect on ridership for the individual stations. Close spacing of stations, however, might also have a negative effect on daily boarding volumes because it increases time for through passengers (Kuby et al., 2004, p.235). A Dutch example of a railroad line where close spacing is pursued is the so-called Lingelijn between the train stations Geldermalsen and Gorinchem. The railway operator is expecting that the opening of several additional stations will yield bigger ridership volumes for the total of the line. Another factor that may account for higher boarding volumes for the same reasons as mentioned above, is the possibility for interline transfers. It has been indicated that train stations that serve as a node with congregating lines from multiple directions generate higher ridership numbers (Kuby et al., 2004, pp.235&239). The possibilities for passengers to easily access an elaborate network through a station, rather than a plain two-direction network, are known as network effects. It improves people's options to access a wider range of destinations and it's thought to have a positive relationship with ridership (Currie et al., 2011, p.548).

Other findings from recent literature are somewhat contrary to the expectations discussed in the previous paragraph in the sense that a train station in areas less densely covered by stations will attract additional ridership through the transit dependent population. Since it Is known that there is always a certain part of the population that is transit dependent (see subsection 3.2.2) it should be considered that stations in certain less covered areas will draw relatively more transit dependent passengers than other stations do (Kuby et al., 2004, pp.235-236). In case of low station density, an individual station's hinterland in larger than in the case of higher densities. In this respect, Kuby et al. (2004, p.235) discriminate between terminal stations and other stations. Terminal stations are train stations at the end of a line, which consequently only offer service in a single direction. Examples of terminal stations in The Netherlands include Vlissingen, Den Helder, Harlingen Haven and Emmen. A dummy variable for terminal stations is shown to be positively related to ridership of light-rail in the United States (Kuby et al., p.239).

3.3.4 Basic train station qualities

Waiting time and transfer time at a train station are experienced negatively by train passengers (Hof & Koopmans, 2006, pp.7-13). Both are nevertheless inherent to travelling by public transportation. A better quality of waiting and transfer time might help to increase a station's ridership. Facilities at a station can help to improve the quality of time spent at a station. Easy station access is frequently mentioned as an important criterion (Currie et al., 2011, p.547; Van der Bijl & Maartens, 2010, p.13). Entrances need to be easy to find and have an airs that is inviting. A station should be well ordered, safe and clean to make the time people spend on their transfer from access journey to train or between trains as comfortable as possible. Commercial facilities can help commuters to put their spare minutes to use. Convenience stores for grocery shopping, food stands for a quick bite or an easy dinner, and small stores to buy gifts, books or magazines enable this (Van der Bijl & Maartens, 2010, pp.12-15). The problem is that almost all commercial facilities at train stations in The Netherlands are exploited by one of NS's subsidiary organizations. This means that the supply side is barely diversified. Stations of similar size and magnitude have almost the same range of facilities.

A higher experienced sense of safety at a station is associated with higher ridership through improved quality of time spent at a train station. The association applies to social safety rather than technical safety of the infrastructure. Social safety is usually subjective, because it is derived from the personal experiences of passengers. An objective subsidiary is the number of registered incidents per station (Hof & Koopmans, 2006, pp.15-19). A problem with this measurement is that it might not adequately depict the apprehended safety and feeling of comfort.

3.4 The built environment

In order to comprehend general travel behavior, it is necessary to analyze the physical structures of the built environment of transit nodes (Cervero & Kockelman, 1997, p.201). Transit ridership is said to be significantly higher in areas that are built according to the principles of transit-oriented development (Sung & Oh, 2011, p.81). Urban environments that adhere to the principles of transit-oriented development (TOD) have a high population density, diverse, mixed land-use, and a pedestrian-friendly design and consequently are more likely to induce the use of public transportation than a sprawling, auto-oriented environment (Cervero & Kockelman, 1997, p.201). Since it is expected that density, diversity, and pedestrian-oriented design will lower the rates of personal car trips (Cervero & Kockelman, 1997, 211), it is hypothesized that these are explaining factors in transit ridership in general and, subsequently, in rail ridership. This section discusses the relevance of the built environment for rail ridership according to the three Ds of transit-oriented development: density, diversity and design.

3.4.1 Density

According to Kuby et al. (2004, p.22), there is an almost unquestionable positive relationship between population density and transit ridership, since density has proven to be a significant determinant for ridership both at a metropolitan and at a station environment level. Other studies have acknowledged the relevance of density as well, stating that transit ridership volumes increase with higher values of residential density (Sung & Oh, 2011, p.76). Currie et al. (2011, p.546) go further by stating that "nearly every study that has focused on transit ridership has provided evidence that density is the primary determinant of transit ridership." Such a positive relationship is probable, because a high density means that a relatively high number of people reside within the direct sphere of influence of stops and stations. All in all and based on recent international literature, it can be concluded that population density can be hypothesized as a major determinant of rail ridership in The Netherlands. Nevertheless, the relevance of density for transit ridership might in actuality be attributable to mixed land-use. Since density and mixed land-use, according to the principles of TOD, frequently co-exist, they could show a certain degree of co linearity (Cervero & Kockelman, 1997, p.202). The following sub-section discusses the influence of mixed land-uses on rail ridership.

3.4.2 Diversity

Diversity, with regards to the theory on TOD, refers to mixed land-use in urban environments. Mixed land-use is the case when a combination of two or more land-use types, such as residential, commercial and business, among others, is found within a single area. Based on this assumption, governments should apply a mixed land-use strategy near transit nodes in order to generate higher ridership volumes (Sung & Oh, 2011, p.81; Verburg et al., 2005, p.27). Mixed use ensures a combination of residents and employees, which generates ridership throughout the area. Determinants for employment that are explicitly mentioned are healthcare, leisure facilities and retail (Verburg et al., 2005, p.22).

Rail station areas that have a large degree of diversification of land-use tend to show higher ridership than areas with single-use functions and areas with low degrees of mixed land-use. This indicates that the TOD principle of diversity is an important factor for ridership generation (Sung & Oh, 2011, pp.80-82). Not all land-use combinations have the same positive effect on ridership though. While train station areas with residential, commercial, business, and other use have significantly higher rail ridership, train station areas with a mix of only commercial and business use do not show higher ridership (Sung & Oh, 2011, p.76). Akayima and Okushima (2009, p.16) however, found that commercial facilities in and of themselves have a positive effect on ridership, especially among older age groups. They have identified a close correlation between commercial development in train station areas and traffic flows. It is nevertheless hypothesized that the mixture of residential and non-residential land-uses, as opposed to the mixture between multiple non-residential uses, is an important determinant of rail ridership (Sung & Oh, 2011, pp.80-81).

3.4.3 Design

The third principle of TOD-based planning, besides density and diversity, is design. In the context of TOD, design refers to the physical build-up of an urban environment, and more specifically to the structure of the street network. Street network and urban design planning approaches are major factors affecting rail ridership (Sung & Oh, 2011, p.81). Several studies have suggested or identified pedestrian zone length in cities as an influence on daily transit boardings (Currie et al., 2011, p.548). Others have identified the density of four-way intersections, the percentage of driveways on streets, and the number of dead-end roads as determinants for ridership (Sung & Oh, 2011, pp.78-82). The latter three factors have an influence modal choice, since they either encourage car usage or walking. For example, a high percentage of driveways on streets enables large single-occupant vehicle traffic flows, while a high number of dead end roads indicates lower car accessibility and is therefore identified with walking. This means that when a train station area has a higher value of driveways, the use of the automobile as the main mode of transportation is encouraged indirectly, which results in a lower share of rail ridership and transit ridership in general. Subsequently, a train station area that also has a higher total length of roads and a higher average building complex surface, in addition to higher driveway values or by itself as a single determinant, tends to have lower rail ridership than others (Sung & Oh, 2011, p.78).

The number of four-way intersections in a train station area is an indicator for the structure of the road network. A very high number indicates that the street network shows the characteristics of a tight-knit Manhattan grid (Cervero & Kockelman, 1997, p.217). Subsequently, it is an indicator for relatively small building complex surfaces. The closer four-way intersections are to each other, the smaller the spaces in between, which consequently leads to smaller building complex areas. A high number of four-way intersections is associated with higher ridership values because it usually involves a pedestrian-friendly, grid type street network. It is shown that this has a statistically significant, positive relationship with rail ridership (Sung & Oh, 2011, p.81). Furthermore, limited on-street parking that abuts retail and other commercial facilities, as it is common in numerous US cities, is also associated with lower personal vehicle miles traveled and might be an indicator for transit ridership (Cervero & Kockelman, 1997, p.217).

3.5 Other determinants for rail ridership

The three previous subsections discriminate between ridership affecting factors by categorizing them under either train users' characteristics (section 3.2), train station and network characteristics (section 3.3), or the built environment (section 3.4). There are, however, other factors affecting rail ridership that are not discussed in either of these sections. The determinants for rail ridership that have not been discussed so far will be presented in this section. All are shortly discussed along with their relevance for the particular case of The Netherlands.

Factors that have been mentioned as being of influence on rail ridership are fares, ticket integration, reliability of service, average train speed, congestion, political support, weather conditions, and the presence of large institutions, facilities or events that can draw large crowds (Kuby et al., 2011, p.547; Currie et al., 2011; Van Vuuren, 2002).

3.5.1 Fares

The first, fares, obviously has an effect on ridership (Currie et al., 2011, p.546). When tickets get more expensive, the demand for train rides will go down. This is basic price elasticity. The influence of ticket pricing is, nevertheless, not easy to determine, because prices are not exactly the same throughout The Netherlands and do not change so often that elasticity can be easily measured (Van Vuuren, 2002, p.72). Train ticket prices for NS operated lines are all based on the same system. The only exceptions are Fyra, which is a relatively new high-speed line connecting Amsterdam, Rotterdam and Breda, and ICE International, which is an international train that can be used for domestic trips between Amsterdam Central station, Utrecht Central station and Arnhem against an additional fee. There is price diversification between different railway operators, to the extent that they base their fares on different systems, which can result in relative differences. This is due to the fact that the national lines are subject to the demands of the national government while the regional lines are let out in contract by the provinces. It can be questioned whether people are aware of the relative price differences between different contractors. For one, fare differences are subtle. There are examples of trips on regional lines that are relatively cheaper, but also which are more expensive. The causes are complex and irrelevant for now, so they will not be further discussed here. Secondly, the use of the OV chipkaart, which is a nationwide system of digitized paying for public transportation, is thought to reduce awareness of fares.

3.5.2 Ticket integration

The second of factors not discussed in the previous sections is ticket integration. This is the concept of being able to travel on different lines, through different regions, and with different modes without the need to purchase multiple tickets. Ticket integration is relevant when different rail systems exist simultaneously. It is achieved when the need to go through a ticket purchase when transferring is absent (Currie et al., 2011, p.548). With the introduction of the OV chipkaart in The Netherlands, ticket integration is more or less achieved. Though options for old-fashioned paper tickets are still existent and multiple check-ins are still required between different modalities and contractors, the OV chipkaart enables users of public transportation to travel without having to purchase separate tickets. Because there are no longer differences in ticket integration in The Netherlands, there seems to be no reason to assume that ticket integration is a differentiating factor for rail ridership in The Netherlands.

3.5.3 Reliability of service

Reliability of service has been named as a factor influencing ridership (Currie et al., 2011, p.547). Reliability of service is the matching degree between scheduled and actual public transport travel times and its time-related impacts on passengers. In an ideal situation, all trains depart from and arrive at a train station according to the schedule. This would mean that the service is 100 percent reliable. In reality however, operations are subject to delays due to disturbances of all kinds, such as bad or unexpected weather, obstructions from other traffic, technical malfunctions or human behavior. Impacts of service reliability on passengers can influence the in-vehicle time, waiting time, and the probability of finding a seat. Unreliability can make people change the route or the moment of their journey, can lead to the cancellation of a journey, or it can make people change their mode of transport. All of the above have a negative impact on ridership, since travelers have strong risk aversion (Van Oort, 2011, pp.19-34; Van Vuuren, 2002, p.116).

Data on reliability of service on the Dutch railroad network is annually provided by the operators. Punctuality of NS is measured by ProRail at the destination station. A train is considered “on time” when it arrives no later than five minutes after it was scheduled. Until recently, trains were considered “on time” up until a three minute delay. That is why data are available for both the 3 minute and the 5 minute benchmark. These data show that service reliability has increased over the last ten years (treinreiziger.nl, 2010). A positive relationship between service reliability of trains in The Netherlands is hypothesized, but since service reliability data per train station is lacking, this is expected to be difficult to affirm.

3.5.4 Average train speed and road congestion

A variable that is affected by service reliability is average speed of operation. Lower reliability of service can results in lower average speed in case of delays. Ridership on route level for light-rail, however, has been shown to be negatively influenced by speed in multiple regression analysis (Currie et al., 2011, p.557). It is expected that the negative relationship is due to the fact that highly urbanized areas have higher ridership, consequent longer dwell times, and shorter stop spacing. This leads to lower average speed (Currie et al., 2011, p.558). For Dutch heavy rail, it is expected that speed in and of itself is not affecting ridership per station. Rather, it is assumed that the total travel time of a journey that includes rail is weighed against total travel time for the same journey with a different mode of transport. That is where congestion comes into play. The speed advantage the automobile might have over journeys that include the train can diminish or even disappear as a result of congestion around the home-end of the trip, along the route, or at the destination-end of the trip (Kuby et al., 2004, p.227).

3.5.5 Political support

Knowles has found that strong policy support for public transport, specifically light-rail, can be a driver for ridership (in: Currie et al., 2011, p.547). The question is what is understood as policy support. Government subsidies for the exploitation of certain lines or for the construction of new stations can be regarded as direct support for public transit. Direct support for public transit can also be direct aid to the specific user groups through discount cards, such as the Dutch public transportation passes for students or reduced fare passes for children and elderly. Indirect support can be the design of public space in such a way that it generates bigger ridership volumes (TOD, see subsection 2.4.3). Financial support from governments is already shown through ticket prices. Indirect policy support is, as said, shown in public space. Still, some subsidized regional heavy rail lines in The Netherlands generate remarkably high ridership volumes. An example is the so-called Valleilijn, which connects the cities of Amersfoort and Ede. It is heavily subsidized by the province of Gelderland. More money is not a direct driver for ridership, but it aids higher frequencies and better quality, which can attract more travelers.

3.5.6 Weather conditions

Research by Kuby et al. (2004) in the United States has shown that climate can have an influence on ridership. Light-rail systems in US cities with moderate climates can expect higher ridership volumes per station than cities that suffer from more extreme temperatures. This can be explained through the comfort experienced while walking to and from stations and while waiting in a moderate climate, as opposed to the discomfort experienced on hot, humid or extremely cold or rainy days (Kuby et al., 2004, p.241). Though there are small differences in temperatures between the coastal regions and the non-coastal regions of The Netherlands, there are no different climates. This indicates that weather and climate do not have a diversifying effect on rail ridership in The Netherlands.

3.5.7 Crowd-drawing attractions

The presence of large institutions, facilities, events, or other large-scale crowd-drawing attractions can increase the use of public transit. Examples are, among others, hospitals, colleges, high schools, museums, zoos, theme parks, fairs, festivals, sports venues and celebrations of national holidays. Fairs, festivals, sports venues and celebrations of holidays only generate additional ridership on the respective days. This means that these days are not representative for average daily ridership. Most theme parks are not open during the winter months. Attractions that are open year-round mostly draw visitors during holidays and on weekends. Colleges and schools do not generate ridership during holidays. Hospitals are expected to be insensitive to holidays and days of the week, though visits might be higher on weekends. Still, the effect of hospitals on rail ridership is expected to be limited. A limited affect is also expected for museums and other year-round attractions. To add 50 daily boardings, an attraction already has to generate (50 * 365 =) 18,250 train-riding visitors annually. Assuming that no more than 10 percent of the visitors travel by train, this would require a total of more than 180,000 annual visitors. Since only five percent of the 773 museums in the Netherlands have more than 100,000 visitors (CBS Statline, 2010), and considering that a 10 percent ratio for visitors who travel by train already is a high estimate, it is assumed that museums and other year-round attractions only have a minor impact, if any, on rail ridership. The effect of high schools is also expected to be minor, since most children in The Netherlands will use the bicycle to get to school. The impact of large colleges and universities is expected to be of more significance. Considering that almost every college enrollment equals the possession of a public transportation student pass, a large college can add significantly to ridership. Research in the United States, however, has shown that college enrollment did not have a significant influence on ridership (Kuby et al., 2004, p.238). An equivalent of the Dutch public transportation student pass is not prevalent, though.

3.6 Conclusion

This chapter discussed a number of factors that are associated with train ridership in recent literature. The relevance of the various factors for the case of The Netherlands and the Dutch railways has been discussed and their respective effects were hypothesized. The chapter aimed at finding drivers of train ridership. In order to do so, a distinction was made between train users' characteristics, station and network characteristics, and the built environment. Generally speaking, train users are characterized by the large amount of time they spend on

Table 3.1: Hypothesized drivers for train ridership in The Netherlands

Category	Ridership driver	Source
Users	Employment	Bakker & Zwaneveld, 2009
		Verburg et al., 2005
	Level of education	Verburg et al., 2005
	College enrollment	Kuby et al., 2004
		Verburg et al., 2005
	Age	Verburg et al., 2005
		Bakker & Zwaneveld, 2009
	Relationship status	Verburg et al., 2005
	Household composition	Bakker & Zwaneveld, 2009
		Verburg et al., 2005
	Degree of urbanization of place of residence	Verburg et al., 2005
	Degree of urbanization of destination	Verburg et al., 2005
	Car ownership	De Beynen de Hoog, 2004
	Driver's license possession	Bakker & Zwaneveld, 2009
Station and network	Frequency of service	Currie et al., 2011
		Hof & Koopmans, 2006
	Number of destinations within time buffer	Currie et al., 2011
		Kuby et al., 2004
	Station accessibility by car	Kuby et al., 2004
		De Vos, 2009
	Parking costs	Kuby et al., 2004
	Number of park-and-ride spaces	Kuby et al., 2004
	Accessibility by bicycle	Givoni & Rietveld, 2007
		Keijer & Rietveld, 2007
	Interconnectivity with other types of public transit	Currie et al., 2011
		Givoni & Rietveld, 2007
		Van der Bijl & Maartens, 2010
	Frequency of bus and light-rail service	Currie et al., 2011
	Possibility for interline transfers	Currie et al., 2011
		Kuby et al., 2004
	Easy station access	Van der Bijl & Maartens, 2010
		Currie et al., 2011
	Quality of waiting time	Hof & Koopmans, 2006
		Van der Bijl & Maartens, 2010
	Commercial facilities at train station	Hof & Koopmans, 2006
	Perceived safety at train station	Van der Bijl & Maartens, 2010
	Station spacing	Kuby et al., 2004
Built environment	High density	Cervero & Kockelman, 1997
		Currie et al., 2011
		Kuby et al., 2004
		Sung & Oh, 2011
	Mixed residential and non-residential use	Akayima & Okushima, 2009
		Sung & Oh, 2011
		Verburg et al., 2005
	Pedestrian-oriented design	Cervero & Kockelman, 1997
		Currie et al., 2011
		Sung & Oh, 2011
Other	Reliability of service	Currie et al., 2011
		Van Oort, 2011
		Van Vuuren, 2002
	Road congestion	Kuby et al., 2004
	Policy support	Currie et al., 2011

work and education. The latter is in accordance with the fact that a quart of all train kilometers is made by students who own a public transportation student pass. Furthermore, there is an overrepresentation of singles and young adults among train users. Public transportation users have, on average, fewer young children than car users. They are further characterized by a highly urbanized place of residence and a highly urbanized destination. The use of public transportation in general and the train in particular, are popular among households with a strong orientation on professional work, while the average level of education is higher for train users than for car users. The transit dependent traveler, or public transit captive, is a traveler who lacks other options for transportation than public transit. This means they either do not own a car or that they cannot use the car they own because it is shared among a multiple-person household.

Other train ridership drivers are related to the railroad network and the train station and with the built environment of the station. Increased frequency of service and a higher number of destinations that can be reached within a shorter time are hypothesized to draw larger crowds. A higher frequency of service reduces average waiting time and, possibly, the average transfer time, which adds to the attraction of journeys by train. Other station and network characteristics that are expected to attract passengers include accessibility by car, bicycle, and other modes of public transit. Accessibility of other modes of public transit refers to the physical accessibility (i.e. is there a possibility to transfer?) and the frequency of service. More frequent service by feedering public transit is expected to have a positive relationship with ridership. The possibility for interline transfers at the train station, as opposed to a single two-direction line, is also associated with higher ridership since it is related with a higher number of possible destinations. Station spacing is another aspect associated with train ridership. Transit ridership is said to be significantly higher in areas that are built according to the principles of transit-oriented development, being high density, mixed land-use (diversity), and pedestrian-friendly design. High density of urban areas implies a relatively high number of people reside within the direct sphere of influence of stops and stations along public transportation networks. The positive relationship that exists between density and transit ridership is widely acknowledged. Factors named in literature as being of influence on ridership that do not directly relate to either train user's characteristics, network and station characteristics, or the built environment have been discussed separately. They include reliability of service, congestion, and political support. The only crowd-drawing attraction that is might have a significant constant effect on daily boardings is college enrollment.

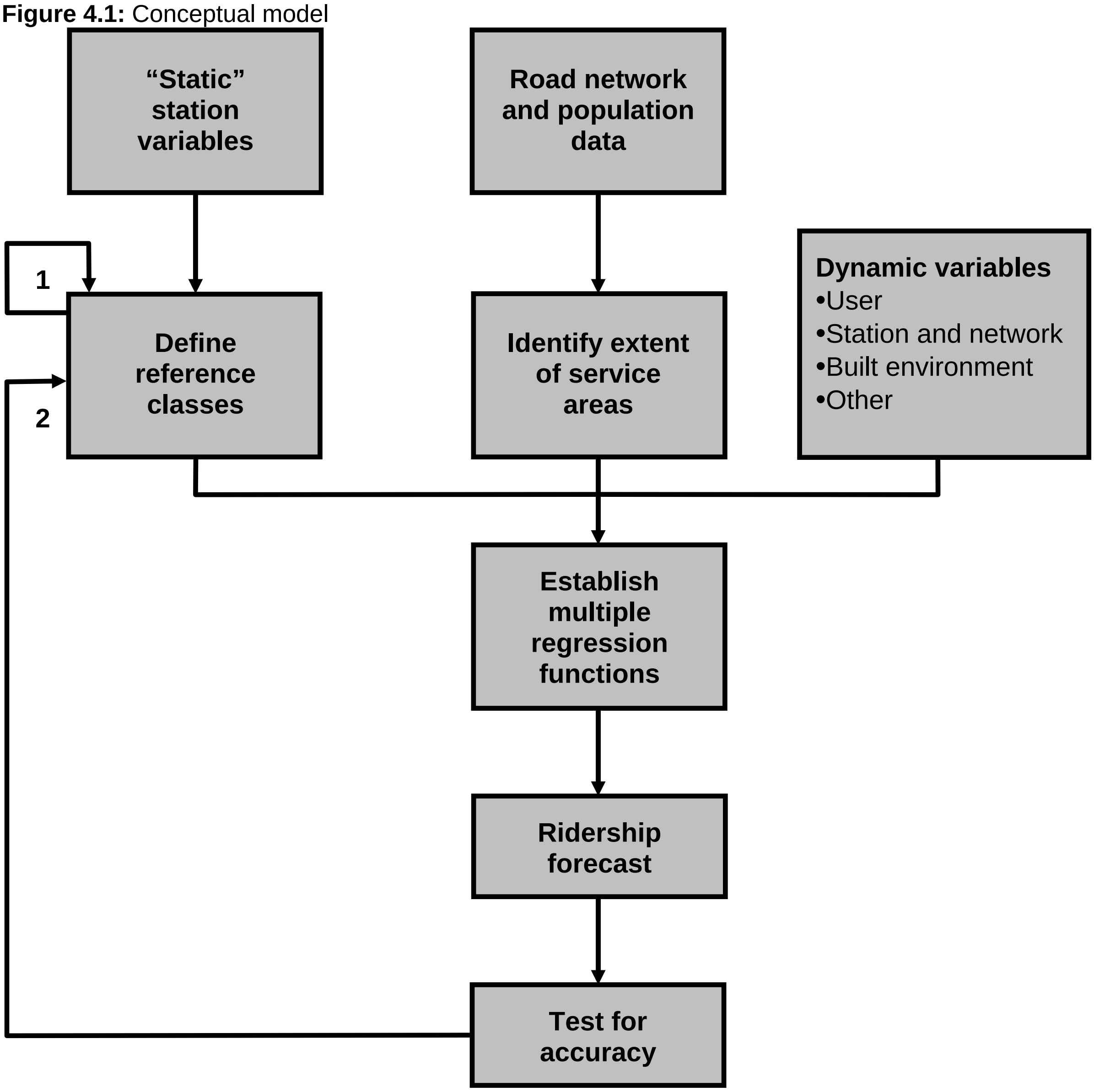
All drivers of train ridership that are hypothesized to significantly affect ridership per station in The Netherlands are included in an overview (see table 3.1). In total, 30 drivers are expected to have a relationship, either positive or negative, with train ridership. All will be tested for significance and impact. The results are presented in chapter 5 and are discussed in chapter 6. The following chapter discusses the methodology applied in this research.

4 Methodology

4.1 Approach

Chapter 3 has discussed the variables that are associated with rail ridership. In this chapter it is discussed how these variables are quantified and what the data sources are, which methods are used to analyze the data, and which procedures and techniques are applied. The methodology of the research roughly constitutes data acquisition, data modification in GIS, definition of the reference classes, and multiple regression analyses in SPSS. These aspects are elaborated on in this chapter. The research procedure followed is depicted in figure 4.1. It shows the defining of reference classes is dependent on iterations. The composition of the reference classes is based so-called static station variables, which are mainly related to the stations' geographical location and their location relative to the town or city they serve. Because the values for these variables typically do not change, it makes them suitable as a basis for reference class discrimination. Furthermore, it is thought that geography matters here because the conditions in the Randstad differ considerably from those outside of it, for example. Also, stations located outside of or on the edge of a town are expected to perform differently than downtown stations.

Figure 4.1: Conceptual model



Iteration 1 in the conceptual model is included because initially, the formation of the reference classes should lead to an optimal balance between the number of cases (*n*) within the reference class and the internal homogeneity of the reference classes. Achieving this balance requires repetition by trial-and-error. The second iteration is included because the reference classes need to be redefined in case they yield inaccurate forecasts. This is related to the aforementioned reference class problem, which is basically the question of which variable or combination of variables should be used to base the reference classes on. Inaccurate forecasts are expected to be a direct result of inapt reference classes used. As a consequence, the definition and formation of the reference classes make up a major step of the research.

4.2 Data

As of January 2012, The Netherlands has 384 train stations in total spread throughout the country. Of these, 64 officially hold the intercity train station status. This means that all passing trains make a stop at the station. These stations are excluded from the analysis, since newly planned stations are hardly ever destined to hold the intercity station status. Of the 320 stations left, only those opened and operated before 2004 are included in the analysis. The stations opened from 2004 through 2008 provide the possibility to check and weigh the generated forecasts for their accuracy, since actual ridership numbers for at least the year 2009 are available for these stations. Stations opened in 2009, 2010, and 2011 are completely left out because data for these stations is either lacking or unavailable. The station in Eijsden is excluded because it has been closed in the mean time. All in all, the population consists of 296 currently operated train stations that are suitable for the multiple regression analyses (see section 4.3). The variables used in for multiple regression are summarized in table 4.1 and are further discussed in the following subsections.

4.2.1 Dependent variable: number of daily boardings

Data for the number of daily boardings per station are available for the years 2004, 2005, and 2006. It has been provided by NS and has been published online. They include both access and egress passengers, but exclude passengers transferring between trains. The numbers are said to be representative for an average weekday. Boardings per station in 2006 ranged from 54 at Enschede De Eschmarke to 39,555 at Amsterdam Sloterdijk. The average number of daily boardings for 296 included stations in 2006 equaled 2,433. This number is higher than the averages for 2004 (2,247) and 2005 (2,356). This is in compliance with the lore that there is a trend of increasing train usage in The Netherlands. Data are neither available for the years prior to 2004 nor for the years following 2006. Exceptions to this are the data for the year 2010, which have become available early 2012, and the stations opened between 2004 and 2008. Ridership numbers for these stations have been published by ProRail (Projectteam Toepassing Norm ProRail, 2009). The number of years that data are available for is dependent on the startup year. Almere Oostvaarders station, which was opened in 2004, has data available for 2005 through 2009. Amsterdam Holendrecht, opened in 2008, has data available for 2009 and 2010 only.

4.2.2 Spatial data

For several purposes of this research, spatial data is required. Firstly, a dataset of all 384 train stations in The Netherlands is required for the calculation of their service areas and to join other data to. Since such a dataset was not readily available for this research, a shapefile has been created containing points representing the train stations. The station points have served as the input for the calculation of the stations' network distance buffers, for which a road network dataset was also required. The road network used in this study was derived from TOP10NL. This was preferred over the Open Street Map dataset because of its consistency. Thirdly, a polygon shapefile containing all 4-digit zip codes in The Netherlands as of 2006 and a polygon shapefile containing all municipalities as of 2006 were required to join the socio-economic data to (see section 4.2.3). Both are available at the Faculty of Geosciences of Utrecht University. BBG was used to derive land use from.

4.2.3 User variables

In total, the literature research yielded ten supposed drivers of train usage. The main source for such data in The Netherlands is CBS Statline. Population data, household data, and car ownership data are provided per 4-digit postal code for the whole of the country for the year 2006. The data on household composition are considered as representatives for relationship statuses. It is assumed that multiple person households are a good indicator for single family households. Degree of urbanization data were acquired through Statline as well, though this is only available per municipality. Unfortunately, data on driver's license possession was not available. Data on the number of jobs were unavailable for postal codes but could be acquired on a municipal level. Data on enrollment at institutions of higher learning were gathered at various sources but they are not subdivided for separate locations. Data at an institutional level are not well applicable, due to the fact that institutions are frequently operating from various locations which might be served by different train stations or might even be in separate cities.

Table 4.1: Independent variables used in the multiple regression analyses

Independent variable	2km	Short name	Minimum value	Average	Maximum value
Population	x	POP_TOTAL	127	13,888.7	71,839
Population of age 15 - 65	x	POP_15T65	86	9,334.3	54,103
Non-western immigrant population	x	POP_ALLOCH	0	1,821.6	33,585
Households	x	HH_TOTAL	52	6,208.4	39,677
Single-person households	x	ONEP_HH	12	2,300.2	22,575
Multiple-person households without children	x	MP_HH_NC	21	1,785.5	7,546
Multiple-person households with children	x	MP_HH_WC	19	2,122.8	9,549
Car ownership	x	CAR_OWN	61	5,532.5	22,103
Intersections	x	CROSSINGS	32	440.5	1,066
Built up percentage	x	BUILTUP_P	1.3	42.4	81.0
Rate Commercial-Industrial-Institutional/Residential	x	CII_RES_RATE	0	0.49	15.6
Total stations accessible within 60 minutes by train		TOT_60	8	30.7	90
Intercity stations accessible within 60 minutes by train		IC_60	0	7.0	24
Stopping stations accessible within 60 minutes by train		STOP_60	3	23.7	69
Total stations accessible within 30 minutes by train		TOT_30	1	7.7	29
Intercity stations accessible within 30 minutes by train		IC_30	0	1.8	9
Stopping stations accessible within 30 minutes by train		STOP_30	0	5.9	20
Frequency of service by train per hour		FREQUENCY_oS	1	6.1	42
Possiblity to take BTM		BTM_BINAIR	0	n.a.	1
BTM frequency of service per hour		BTM_TOTAL	0	12.0	118.5
Additional, partial intercity service		TYPE_SCORE	0	n.a.	1
More than two directions of service		TRANSFER	0	n.a.	1
Employment within municipality		JOBS_MUN	1,110	43,455	512,080
Average household income within municipality		INCOME_AV	11,100	13,224	22,300
Degree of urbanization 1		URBANIZATION_1	0	n.a.	1
Degree of urbanization 2		URBANIZATION_2	0	n.a.	1
Degree of urbanization 3		URBANIZATION_3	0	n.a.	1
Degree of urbanization 4		URBANIZATION_4	0	n.a.	1
Randstad		RANDSTAD	0	n.a.	1
Station category 5		CATEGORY_5	0	n.a.	1
Station category 6		CATEGORY_6	0	n.a.	1

Source: own adaptation, 2012

4.2.4 Station and network variables

The number of train stations accessible per station within a certain time frame is derived from a rail journey time matrix provided by Goudappel Coffeng. The matrix contains the time it takes to travel from one station to every other station in The Netherlands and consequently contains all stations operated in 2008. It was chosen to apply time frames of 30 minutes and 60 minutes because these seem logical boundaries. Furthermore, we diversified between intercity train stations and other train stations. This has resulted in six variables, being the number of intercity stations accessible within 30 and 60 minutes, the number of other stations accessible within 30 and 60 minutes, and the totals within 30 and 60 minutes. These are also considered to be indicators for station spacing. A high number of stations accessible within a time frame is thought to indicate small spacing and, conversely, a low number of stations accessible equals large spacing. The frequency of service per station was provided by Goudappel Coffeng. The data are per hour during morning rush hours and cover the year 2008. For practical purposes, it was assumed that these frequencies were equal to those in 2006. The frequency of bus, light rail, and subway services per station were provided by Goudappel Coffeng as well. Here too, the data refer to the year 2008. No distinction is made between bus, light rail, or subway, so the data contain only the total of stops made by other types of public transportation at the station. Again, this is per hour during morning rush hours. A dummy variable was created for the feature of interconnectivity with bus, light-rail, and subway where 0 equals no interconnectivity and 1 equals interconnectivity, regardless the frequency. Another dummy variable was created for the possibility for interline transfer, for stations like Elst, Geldermalsen, and Weesp, where 1 represents the possibility to transfer.

Accessibility of stations is measured by the number of 3-way and 4-way intersections within a two kilometer network buffer. These numbers were acquired in GIS. By buffering the intersections' centerpoints, it was possible to create single polygons for intersections with lanes represented by multiple centerlines and for roundabouts. The application of 10 meter crow-flight buffers lead to overlapping centerpoint buffers at large intersections and roundabouts (see section 4.5). Consequent dissolving of overlapping buffers made it possible to count the actual number of intersections within the stations' service areas. No diversification has been made between car accessibility and slow traffic accessibility because the road network was missing specific attribute values.

Data on the number of park-and-ride spaces and the parking costs were gathered from several sources, but since it was not possible to acquire data for all 297 stations, it was chosen to exclude these from the analyses. The commercial facilities variable was excluded because of the absence of a source of data. It seemed impossible, given the available time, to create these data. The perceived safety at a station and the quality of the waiting time are subjective matters which are difficult to measure and quantify. Because apt data could not be found, both are excluded.

4.2.5 Built environment variables

Ideally, density is represented by the number of addresses for a standardized area. Such data could, however, not be obtained. The number of households per 4-digit zip code is available, but that is insufficient due to the missing commercial addresses. Another alternative is the degree of urbanization of the municipalities, available at CBS Statline, on a scale from 1, being highly urbanized, through 5, being highly rural. This scale is based on the address density, but its use in this research is limited because it does not diversify between stations within the same municipality. Furthermore, the degree of urbanization is not as detailed as one would ideally strive for in a multiple regression analysis. Because of its ordinal scale, it requires a dummy variable for each value in order for it to be properly incorporated into a regression analysis. Due to these limitations, the percentage of the stations' service areas that are built upon with residential, commercial, industrial, or institutional are added. These percentages have a higher level of detail and can easily be obtained from land use types in BBG. The downside of using BBG here is that a single large dwelling can account for an area equal to that of an apartment block including dozens of addresses or that one business with a few terminals can take up as much space as several others combined.

Diversity of the built environment is measured by the ratio between residential land use on the one hand and commercial, industrial, and institutional land use on the other. Both are calculated in GIS in square meters within the stations' service areas. The ratio is calculated by dividing the sum of the area of commercial, industrial, and institutional by the area of residential. A rate of 1.00 thus means that both are equally prevalent. A rate under 1.00 means that residential is the dominant land use type. When the rate tops 1.00, commercial, industrial, and institutional land use combined are more prevalent than residential land. Land use data derived from BBG have been used as input for these calculations. The third variable from the built environment category, pedestrian-oriented design, has been projected by the number of intersections in previous studies, such as Sung and Oh (2011). In this case, the number of intersections is also used as an indicator for station accessibility (see section 4.2.4).

4.3 Multiple regression

With a multiple regression function, one tries to explain the relationship between a dependent variable Y and multiple explanatory variables. In most cases, the explanatory variables not only correlate with the dependent variable, but also among each other. With multiple regression, one can account for these correlations while calculating a linear function. Such a function, with *n* explanatory variables, looks like the function below.

$$Y = a + b_1 X_1 + b_2 X_2 + ... + b_n X_n$$

Here, Y is the predicted value for the dependent variable, which is the number of daily boardings per station in this case. The intercept is represented by *a*. The intercept is the point where the function intersects the y-axis. The *b* in this function is the partial regression coefficient. Its shows the change in Y caused by a single explanatory variable while all other variables are kept constant. It is positive when there is a positive relation between the explanatory variable and the dependent variable.

The purpose of multiple regression is to identify the explanatory variables for the daily ridership volumes of Dutch train stations. Secondly, these variables should lead to an explained variance as high as possible or, in other words, minimize the residuals. Thirdly, multiple regression is suitable for eliminating the effect of other explanatory variables, as opposed to simple linear regression, where there is only one explanatory variable X.

4.3.1 Prerequisites

There are five prerequisites one has to consider when applying multiple regression. Firstly, the variables have to be measured in an interval or ratio scale. It is possible to include categorical variables, but they need to be converted into dummy variables, which are in a dichotomous scale. The general rule is that the number of dummy variables required for a single categorical variable equals the number of categories minus one. Secondly, all explanatory variables will have to have a theoretical causal relation with the dependent variable Y. The theoretical bases for causal relations between the identified explanatory variables and the dependent variable have been elucidated in chapter 3. Thirdly, the model requires linearity. Subsequently, the residuals are normally distributed and are independent from X. The latter attribute is known as homoscedasticity. The prerequisites regarding linearity and residuals can be verified by using SPSS.

The fifth prerequisite for multiple regression is that multicollinearity is non-existent. This means that the explanatory variables do not correlate too much among each other. Correlation is measured by the correlation coefficient, or Pearson product-moment correlation coefficient, and is represented by *r*. It is solely a measure for correlation. As long as $-0.9 \geq r \leq 0.9$, the respective two explanatory variables can both be included into the analysis. When *r* passes ±0.9, one of the two should be left out. In case of the latter, it means that the two variables practically account for the same part of the variance in Y.

More information on multiple regression and its prerequisites can be found in De Vocht (2005a, pp.145-156).

4.3.2 Multiple regression in SPSS

In SPSS, one can choose from five types of methods for performing a multiple regression analysis. The standard method Enter adds all selected explanatory variables into the model simultaneously. The method Stepwise is different in the sense that it adds explanatory variables into the model one by one in accordance with their relative importance for the dependent variable. The other three methods, being Forward, Remove, and Backward, are similar to stepwise. The stepwise method is often used because it only includes the significant variables into the model and because it directly shows the relative importance for each single explanatory variable. Furthermore, the stepwise method automatically leaves out one of two heavily correlating variables. This is due to insignificance of the second variable once the first variable has entered the model.

Though a checkup for multicollinearity is not always necessary, it has been chosen to do so anyway in this research. By drawing a correlation matrix, one gains insight in the singular significance of the explanatory variables to the dependent variable. Furthermore, it shows which two variables share an *r* over 0.9. When one has this knowledge before calculation of the multiple regression model, one can choose himself which of the multicollinear variables should be entered into the model. The added value of this knowledge is best explained by an example. When the variables "POP_TOTAL" and "CAR_OWN" have an *r* of 0.94, while the significance of the latter is a little higher than that of the former, the latter could be included into the model at the expense of the former. Both intuitively and theoretically, this seems odd. The number of cars would, in this example, serve as a replacement for the number of people. Theoretically, however, high car ownership should lead to lower numbers of trips by public transportation. In such cases, one can prevent this by leaving the car ownership variable out of the analysis at the cost of the population variable.

It has been chosen to leave extreme outliers out of the eventual models. Outliers can be detected through Casewise Diagnostics in SPSS. In this research, the so-called Zresidual is set at two standard residuals. This is in compliance with the standard normal distribution of the residuals. This means that 95 percent of the cases should be within two standard deviations of the regression line. A large number of cases with a residue of more than two standard deviations indicates that there could be another, unidentified variable accounting for the unexplained variance. When it is shown that the model includes cases with residuals equal to more than four standard deviations, the calculation of the multiple regression function is performed again, excluding the extreme outliers with a Zresidual of over 4 in order to achieve higher accuracy. More information on multiple regression with SPSS can be found in De Vocht (2005b, pp.211-230).

4.4 Reference classes

Theoretically, it is possible to achieve that 100 percent variance of the dependent variable with multiple regression is accounted for. Practically, however, this can usually not be achieved due to various reasons, such as unavailability of data, impracticability of measurement of variables, and unawareness of relationships. This results in a residual, which is the portion of the dependent variable that is not accounted for. To reduce the residual without having to include additional variables, one can choose to apply reference classes (see section 2.5). Instead of calculating a regression function based on the whole population, a regression function is then based on all cases with a certain characteristic or a set of characteristics. Aptly chosen reference classes can result in smaller residuals, but because they are based on fewer cases, it is statistically less reliable. Therefore, a balance between the size of the reference classes and the size of the residuals needs to be sought.

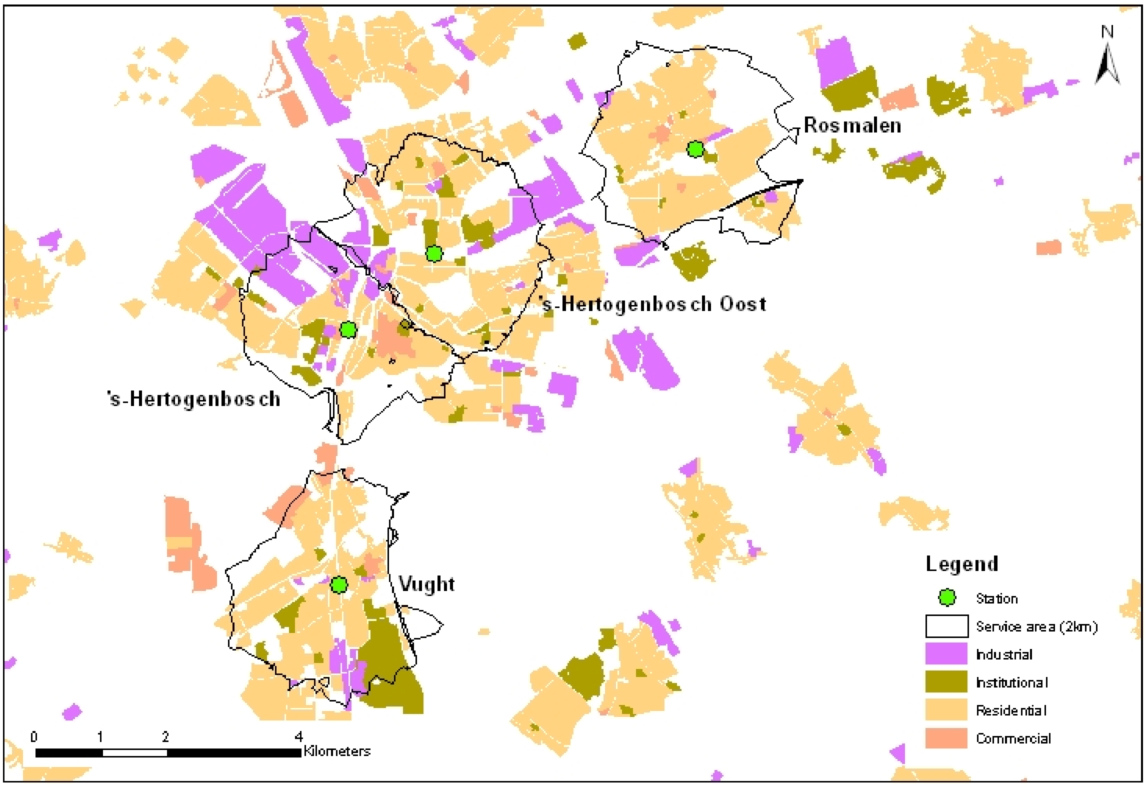
For this research it has been chosen to compose reference classes on static, nominal variables. These are station variables whose values are regarded to stay intact due to geographical location. For the train stations considered in this research, these variables are whether it is within or outside of De Randstad, the station's relative location to the city or town it serves, or a combination of both. Furthermore, customized reference classes are applied to the stations used for validation to find which method yields the most accurate results (see section 4.6).

4.5 Application of GIS

As discussed is subsection 4.2.2 and subsection 4.2.5, GIS plays an important part in the phase of data acquisition and preparation. The software package ArcGIS 10 is used here. Aside from the common functionality, such as Dissolve, Intersect, and Join, the Network Analyst extension has a major role. The Network Analyst provides the user with the means to compute buffers based on actual distance over a network of roads rather than Euclidean distance. In this research, it is used to compute service areas for train stations. The service areas are subsequently used to determine, among others, the number of residents within the stations' spheres of influence. The extent of the service areas is determined at 2000 meter. An analysis of service areas of 500, 1000, 1500, 2000, 3500, and 5000 meter has shown that the 2000 meter network distance buffer's number of residents has the most significant influence on the ridership number, while others are either insignificant or less significant than the 2000 meter service area.

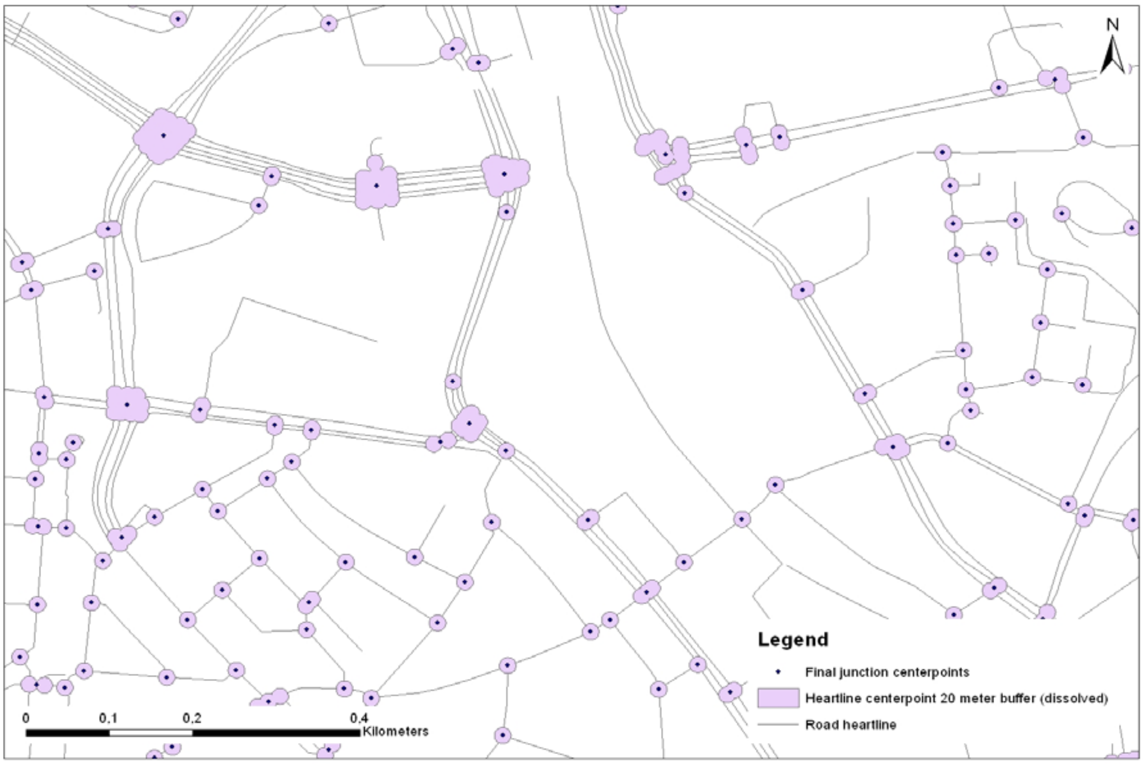
Although intercity train stations are not included into the multiple regression analyses, they are used in the process of calculation of the service areas. It is assumed that people who travel by train choose to access it at the station closest to their home and egress closest to their destination. In cases with two stations within a four kilometer distance of each other, there will be overlap of their respective service areas, unless they are stopped at the point where the two meet. The case of 's-Hertogenbosch and 's-Hertogenbosch (figure 4.2) shows the practical implication. The computed service areas are applied to variables other than the number of residents too. Figure 4.2 depicts the four land use types that are used for the calculation of the diversity variable "Land_use_mix". The cumulative areas of commercial, institutional, and industrial within each station service area are set off against residential land use to calculate the ratio between the two. The ratios are subsequently added to the train station dataset as a separate variable.

Figure 4.2: Land use within train stations' service areas, the case of 's-Hertogenbosch



Source: BBG, 2008.

Figure 4.3: Calculation of the number of junctions based on road heartline centerpoints



Source: TOP10NL, 2008.

Another variable associated with rail ridership at a station level is the number of intersections in the station's vicinity. This cannot simply be derived from the centerpoints of a road dataset because this will yield more points than there are actual intersections. As shown by figure 4.3, certain roads are depicted by multiple lines. Where they are crossed by other road line segments, multiple points will show. To reduce these numbers to one, 10 meter Euclidean distance buffers are drawn around each point. Intersecting buffers are dissolved into single polygons to subsequently depict a single point for each intersection by generating the centerpoint of each polygon emerged from the dissolve operation (see figure 4.3). The buffer distance of 10 meters (radius; the diameter is 20 meters) is chosen because it is both large enough to enable overlapping of all buffers at an intersection, and small enough to avoid overlapping of buffers of different intersections. A brief inspection of the results gave that errors are limited. The results are thus used as a variable in the multiple regression analyses. Unfortunately, no diversification between 3-way and 4-way intersections has been made because a valance renderer in GIS was not available. Diversification between the two can be useful because the latter is associated with higher accessibility than the former.

4.6 Validation

The accuracy of ridership forecasts cannot be fully determined until actual ridership numbers become available for the station at hand. Several years may pass between an initial forecast and the construction and, finally, the opening of the station. Even then, the accuracy of the predicted ridership number cannot be fully determined, since actual ridership numbers may vary within the first few years of operation. This is shown by table 4.2, which compares the generated ridership volumes for stations opened since 2004. Early fluctuations occur until the station has fully rooted in the community. Developments of generated ridership volumes in the first years of stations' existence indicate that it may take four to five years until ridership numbers stabilize.

Table 4.2: Train stations opened since 2004

Station	Established	Ridership '09	Ridership '10	Change	% Change
Almere Oostvaarders	2004	3439	3514	75	2,18
Den Haag Ypenburg	2005	1317	1379	62	4,71
Arnhem Zuid	2005	1945	2109	164	8,43
Voorst-Empe	2006	288	316	28	9,72
Twello	2006	1330	1460	130	9,77
Helmond Brandevoort	2006	833	971	138	16,57
Gaanderen	2006	339	339	0	0,00
Apeldoorn Osseveld	2006	773	967	194	25,10
Apeldoorn De Maten	2006	636	619	-17	-2,67
Amersfoort Vathorst	2006	1633	2364	731	44,76
Utrecht Zuilen	2007	1397	1320	-77	-5,51
Tiel Passewaaij	2007	1230	1152	-78	-6,34
Purmerend Wijdevenne	2007	1578	1269	-309	-19,58
Heerlen De Kissel	2007	419	371	-48	-11,46
Groningen Europapark	2007	862	989	127	14,73
Eygelshoven Markt	2007	149	285	136	91,28
Amsterdam Holendrecht	2008	3111	3024	-87	-2,80
Sum		**21279**	**22448**	**1169**	**5,49**
Mook Molenhoek	2009	x	1224		
Amsterdam Science Park	2009	x	1069		
Maarheeze	2010	x	874		
Heerlen Woonboulevard	2010	x	85		

Source: Projectteam Toepassing Norm ProRail, 2009; ProRail, 2012

The train stations in table 4.1 are used to assess the accuracy of the forecasts extrapolated from the multiple regression function. In order to do so, the research constitutes of using the data for 2006 as input for the multiple regression analyses. Basically, forecasts for the seventeen (plus four) stations in table 4.1 are done as if it were the year 2006. This makes it possible to compare the forecasts against the generated ridership volumes of 2009 and 2010, which are shown in the third and fourth column of table 4.2 respectively. The interpretations of the comparisons will take into account the possibility of fluctuations in passenger flows within the first few years of operation and the fact that it takes some years for a station to come to its full potential. Given that the vast majority of stations in table 4.2 is opened in either 2006 or 2007, it is expected that most stations have not yet grown to their full potential. This should result in forecasts that are on average higher than the generated numbers. The average overestimate of 2010 will be lower than that 2009, since the stations were closer to their full potential in 2010 than they were in 2009.

A disadvantage of using the year 2006 as a starting point is that the results need to be interpreted with care, as discussed in the previous paragraph. It is nevertheless chosen to do so, primarily because of one major reason, being data availability. As discussed in section 4.2, ridership data are unavailable for the years prior to 2004 and the years 2007 through 2009. Data for 2010 have only become available during the late phase of the research. Because important geographical datasets are only available for the year 2008, 2006 was chosen as a starting point rather than 2004 or 2005. The situating depicted in the 2008 geographical datasets is expected to resemble the situation of 2006 more closely than those of 2004 and 2005. Intuitively, it can seem contradictory to still include stations opened in 2004 and 2005 in the list of stations used for validation. Because it takes several years for a station to grow to stabilized passenger flows however, the use of these stations is justified.

Testing of the accuracy of the projected ridership forecasts is the final step of the research. The results of the different reference classes are compared in order to identify the reference class which yields the most accurate forecasts. The results of the research are presented in the following chapter.

5 Results

5.1 Introduction
Recent literature identifies numerous supposed drivers of rail ridership. These have been discussed elaborately in chapter 3. A part of this research is to find which of those variables are and, to which extent they are, of influence on rail ridership at the station level in The Netherlands. That is ventilated in this chapter. The following sections present the results of the multiple regression analyses performed for the various reference classes. A separate section is devoted to each reference class. Each provides details regarding the steps that have been taken. The regression functions and other relevant features are presented as well. Firstly, however, in the following paragraph it is shown why a two kilometer distance service area is applied to all stations in the analyses.

Since the literature research does not provide a decisive answer to the question of which extent to use to determine the stations' service areas, it is included as a research step in and of itself to do so (see figure 4.1 and section 4.1). The goal is to determine which distance threshold needs to be used for the stations' service areas, which are used to calculate the numbers that belong to the variables that have to do with the stations environment and the user characteristics. With simple linear regression, it is found that the number of residents within a distance of two kilometer of the station has the highest influence on ridership, compared to distances of 500, 1000, 1500, 3500, and 5000 meter. The coefficient of determination R^2 of the residents within the 2000 meter network buffer is 0.32, which is higher than 0.31 and 0.27 of the 3500 and 5000 meter buffers respectively. The smaller buffers have coefficients of determination of 0.25 and lower (see appendix 1). The calculation of the coefficients of determination is based on 243 train stations with no more than 6000 daily passengers, with an average of 1672.

5.2 No division

5.2.1 All Stations reference class
The first reference class used in a multiple regression analysis consists of all stop train stations (see appendix 2). The first two MR runs both yielded outliers with standardized residuals of more than 0.4 (see Table A2.3 and table A2.4 of appendix 2). The outliers include, but are not limited to, Amsterdam Sloterdijk, Rotterdam Blaak, and Schiedam Centrum. These were removed and a new attempt was subsequently made. The third run did not yield outliers. Apart from the eight statistical outliers, all stations served as input for the eventual function, which is presented below.

Y = -1260.1 + 247.4 * FREQUENCY_oS + 2.1 * CROSSINGS + 42.2 * BTM_TOTAL -1259.4 * CATEGORY_5 + 101.6 * IC_60 + 650.8 * TYPE_SCORE + 20.7 * BUILTUP_P + 1218.1 * TRANSFER -0.085 * ONEP_HH -347.3 * URBANIZATION_4 -68.6 * TOT_30

With $n = 291$
df = 276
$R^2 = 0.770$
Adj. $R^2 = 0.761$

The model is significant at the 0.000 percent level and explains 77 percent of the variance of Y. Although the Stepwise method for entering variables, which is used here, ensures exclusion of at least one of two variables with mutual multicollinearity, a check for multicollinearity has been performed prior to the analysis. High correlation values led to the exclusion of seven variables. These include five population variables and the station accessibility variables TOT_60 and STOP_30 (see table A2.1 and table A2.2 in appendix 2).

The model meets the criteria regarding linearity and homoscedasticity.

5.2.2 All Stations forecasts
The regression function belonging to the All Stations reference class, as presented in the previous subsection, is used to determine the forecasts for the 17 assessment stations in order to assess the results. The outcomes are presented in table 5.1. The average deviations (positive or negative) of the forecasts as opposed to the actual ridership numbers for 2009 and 2010 are 842 and 799 respectively. The average ratio is 313 for 2009 and 245

for 2010. This means that the forecasts are generally higher than the actual ridership numbers. The highest difference between forecast and generated numbers is found for the case of Arnhem Zuid, with a plus of over 2000 for both years considered. The most accurate forecasts per year are for Gaanderen (plus 55 in 2009) and Purmerend Wijdevenne (negative 6 in 2010). Forecasts for three stations are negative.

Table 5.1: Forecasts from the All Stations reference class

Station	N2009	N2010	Forecast	Deviation '09	Deviation '10	Ratio '09	Ratio '10
Almere Oostvaarders	3439	3514	4164	725	650	725	650
Den Haag Ypenburg	1317	1379	2455	1138	1076	1138	1076
Arnhem Zuid	1945	2109	4290	2345	2181	2345	2181
Voorst-Empe	288	316	41	247	275	-247	-275
Twello	1330	1460	1786	456	326	456	326
Helmond Brandevoort	833	971	-348	1181	1319	-1181	-1319
Gaanderen	339	339	394	55	54	55	54
Apeldoorn Osseveld	773	967	1867	1094	900	1094	900
Apeldoorn De Maten	636	619	1459	823	840	823	840
Amersfoort Vathorst	1633	2364	2136	503	228	503	-228
Utrecht Zuilen	1397	1320	2400	1003	1079	1003	1079
Tiel Passewaaij	1230	1152	-255	1485	1407	-1485	-1407
Purmerend Wijdevenne	1578	1269	1263	315	6	-315	-6
Heerlen De Kissel	419	371	201	218	170	-218	-170
Groningen Europapark	862	989	562	300	426	-300	-426
Eygelshoven Markt	149	285	-601	750	886	-750	-886
Amsterdam Holendrecht	3111	3024	4791	1680	1767	1680	1767
AVERAGE				**842**	**799**	**313**	**245**

5.3 Randstad division

5.3.1 Randstad reference class

A division between the train stations based on whether they are located within or outside of the Randstad depends on the definition used for the Randstad. Here, the whole of the province of Zuid-Holland is included, as well as the southern half of the province of Noord-Holland, with Zaandam as the northern boundary. Furthermore, the Utrecht province stations west of the line Maarssen-Utrecht-Houten are included, as are the stations within the municipality of Almere in the province of Flevoland. Although different definitions of the extent of the Randstad exist, this division seemed reasonable because it is in accordance with the most urbanized areas of western Netherlands.

The stations within the boundaries of the Randstad, as described, were used to compile the Randstad reference class and such served as input for the following function:

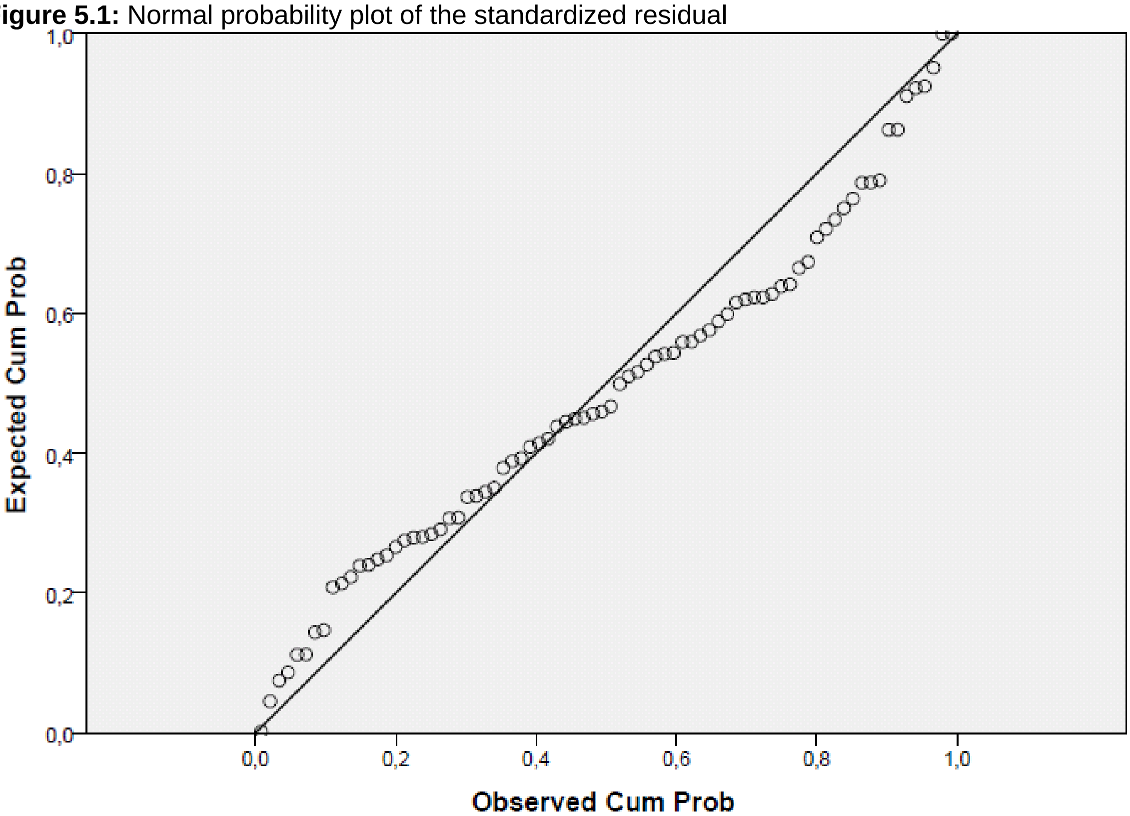
Y = -3624.6 + 745.7 * FREQUENCY_oS + 2490.6 * BTM_BINAIR -1937.6 * CATEGORY_5 + 0.005 * JOBS_MUN

With n = 79
df = 74
R^2 = 0.736
Adj. R^2 = 0.722

Nine variables were excluded due to multicollinearity (see table A3.1 in appendix 3). No cases with standardized residuals of more than 0.4 were identified (see appendix 3 for details). The function does not include the variable POP_TOTAL, though based on theory, and intuitively, it is reasonable to expect it to be of influence on ridership. Adding it to the model with the Enter method, however, does not improve the model and is thus not proceeded. The accepted model is fully significant.

The prerequisites regarding linearity and homoscedasticity are met. Although the dots representing the standardized residuals form a relatively curvy line along the diagonal in the normal probability plot (see figure 5.1), the standardized residuals can still be considered to be normally distributed.

Figure 5.1: Normal probability plot of the standardized residual



5.3.2 Non Randstad reference class

The Non Randstad reference class is made up of the 217 stations located outside of the Randstad as described in the previous subsection. Eight variables were excluded due to multicollinearity (see table A4.1 in appendix 4). Among these was the variable CROSSINGS, which shared a Pearson correlation coefficient of more than 0.9 with POP_TOTAL. The first two MR runs yielded one and two outliers respectively, being Duiven with a value of over 11.0, and Culemborg and Heerhugowaard in the second attempt. These were removed for the third run, which did not show any outliers (see table A4.5 in appendix 4). It resulted in the following function:

Y = -592.8 + 0.101 * POP_TOTAL + 214.0 * FREQUENCY_oS -1040.1 * CATEGORY_5 + 44.4 * BTM_TOTAL -4194.2 * URBANIZATION_1 + 373.2 * TYPE_SCORE + 53.8 * IC_60 -62.0 * TOT_30

With n = 214
df = 202
R^2 = 0.720
Adj. R^2 = 0.709

The model as a whole is significant at the 0.000 percent level. The different variables in the model are significant at, at least, the 95 percent level.

5.3.3 Forecasts from the Randstad and Non Randstad reference classes

Table 5.2 contains the forecasts derived from the regression functions for the Randstad and Non Randstad reference classes as presented in subsection 5.3.1 and subsection 5.3.2. The average deviation for the division as a whole is 1352 for 2009 and 1378 for 2010. The ratios are 402 and 333 for 2009 and 2010 respectively. The deviations for the Randstad reference class are much larger, with averages of 2565 and 2171 for the two years considered. The ratios are equal because the four stations in this reference class all show a plus. When looked at the Non Randstad reference class, one can state that the forecasts are much more accurate, with the exception of the case of Groningen Europapark. The average deviation for the cases remaining after Groningen

Europapark's exclusion for the year 2009 is 625 and for 2010 it is 647. The ratios are 90 for 2009 and 8 for 2010. This means that the model is not overoptimistic (see chapter 6 for further discussion of the results). The most accurate results per year are for Gaanderen in 2009, with a plus of 194, and for Purmerend Wijdevenne with a plus of 109 for 2010. The model resulted in three negative forecasts. Besides for Groningen Europapark, these are for Helmond Brandevoort and Eygelshoven Markt.

Table 5.2: Forecasts from the Randstad and Non Randstad reference classes

Station	N2009	N2010	Randst.	Non Randst.	Deviation '09	Deviation '10	Ratio '09	Ratio '10
Almere Oostvaarders	3439	3514	5804		2365	2290	2365	2290
Den Haag Ypenburg	1317	1379	4227		2910	2848	2910	2848
Arnhem Zuid	1945	2109		3969	2024	1860	2024	1860
Voorst-Empe	288	316		586	298	270	298	270
Twello	1330	1460		1729	399	269	399	269
Helmond Brandevoort	833	971		-357	1190	1328	-1190	-1328
Gaanderen	339	339		533	194	194	194	194
Apeldoorn Osseveld	773	967		1759	986	792	986	792
Apeldoorn De Maten	636	619		1383	747	764	747	764
Amersfoort Vathorst	1633	2364		1249	384	1115	-384	-1115
Utrecht Zuilen	1397	1320	4038		2641	2718	2641	2718
Tiel Passewaaij	1230	1152		356	874	795	-874	-795
Purmerend Wijdevenne	1578	1269		1378	200	109	-200	109
Heerlen De Kissel	419	371		133	286	239	-286	-239
Groningen Europapark	862	989		-3738	4600	4727	-4600	-4727
Eygelshoven Markt	149	285		-391	540	676	-540	-676
Amsterdam Holendrecht	3111	3024	5454		2343	2430	2343	2430
AVERAGE					**1352**	**1378**	**402**	**333**
RANDSTAD					**2565**	**2571**	**2565**	**2571**
NON RANDSTAD*					**625**	**647**	**90**	**8**

* Excluding the case of Groningen Europapark

5.4 Station type division

This section presents the results from the multiple regression analyses applied to the reference classes based on the station types. The three types distinguished here are adopted from a division used by NS based on the stations' locations relative to the city they serve. A station in category 4 is located within the residential area of the city or town it serves and is the only of the major station within the city. Examples include Maarssen, Houten, Geldrop, Zaltbommel, and Anna Pauwlona. Category 5 stations include Tilburg Reeshof, Arnhem Velperpoort, and Amsterdam Sloterdijk. These are usually located outside the city center and serve as a subsidiary to a larger, often intercity train station. Stations that are located at the outskirts of a city or outside of a city, such as in between two settlements, are categorized 6. These include Geldermalsen, Beek-Elsloo, Lichtenvoorde-Groenlo, and Lage Zwaluwe.

5.4.1 Category 4 reference class

Variables excluded due to multicollinearity include IC_60 and STOP_60, STOP_30, POP_15T65, HH_TOTAL, MP_HH_NC, MP_HH_WC, and CAR_OWN (see table A5.1 and table A5.2 in appendix 5). Obviously, there is no use in including the categorical variables, since all stations stem from the same category. The first MR run of category 4 stations showed Duiven station as an outlier. After it was removed, the second run did not yield cases outside of the threshold value (see appendix 5). It resulted in the following function:

Y = -2074.1 + 230.4 * FREQUENCY_oS + 0.115 * POP_TOTAL + 51.7 * BTM_TOTAL + 1204.4 * TYPE_SCORE + 36.4 * TOT_60 + 1027.2 * TRANSFER

With n = 123
df = 116
R^2 = 0.781
Adj. R^2 = 0.769

All variables are significant at the 95 percent level. The model as such is fully significant. The model meets the criteria regarding linearity and homoscedasticity.

5.4.2 Category 5 reference class

Seven variables were removed due to multicollinearity prior to the first run (see appendix 6). The first three runs all yielded a single outlier. In order of appearance, they include Amsterdam Sloterdijk, Schiedam Nieuwland, and Rotterdam Blaak. The fourth run did not, and resulted in the following function:

Y = -1130.1 + 325.5 * IC_30 + 28.7 * BTM_TOTAL + 95.0 * IC_60 + 24.1 * BUILTUP_P

With $n = 84$
df = 79
$R^2 = 0.722$
Adj. $R^2 = 0.708$

All variables are significant at the 95 percent level. The model is significant at the 0.000 level. The model meets the criteria regarding linearity and homoscedasticity.

5.4.3 Category 6 reference class

Seven variables were excluded due to multicollinearity (see table A7.1 and table A7.2 in appendix 7). The variables CATEGORY_5 and CATEGORY_6 did not need to be included because all cases score the same. TYPE_SCORE was excluded for the same reason. The first run did not bring forward any outliers, and the resulting model was thus accepted:

Y = -877.5 + 0.130 * POP_TOTAL + 317.8 * FREQUENCY_oS -109.1 * STOP_30 + 52.6 * BTM_TOTAL + 473.6 * URBANIZATION_3

With $n = 82$
df = 76
$R^2 = 0.786$
Adj. $R^2 = 0.772$

All variables are significant at the 95 percent level. The variable POP_TOTAL is clearly the most important variable in this model, judging by the Beta coefficient (see table A7.4 in appendix 7). The model is significant and meets the prerequisites.

5.4.4 Forecasts for the three Category reference classes

In this section, the forecasts for the 17 assessment stations based on the station category division are presented. The forecasts are included in table 5.3, which contains a separate column for each of the three reference classes. The average deviation for this division is 704 for 2009 and 744 for 2010. The ratios for these years are 378 and 309 respectively. The average deviations for the Category 4 reference class are based on the only two cases among the 17 stations that fit within the class, being Twello and Gaanderen. Both have generated more passengers in 2009 and 2010 than forecast, which results in negative ratios for both years.

Of the 17 assessment stations, 14 are category 5. The average deviation for this class is 806 for 2009 and 848 for 2010. These numbers are higher than the averages for the division as a whole. The corresponding ratios are higher too, with 468 for 2009 and 396 for 2010.The only case matching category 6 is Voorst-Empe. Not surprisingly, the average for the class equals the values for this individual station. The generated ridership numbers for both 2009 and 2010 are lower than forecast.

The only station with a negative forecast in this division is Tiel Passewaaij, while it generated about 1200 daily passengers in both years considered. In thus has significant influence on the results. The station with the highest deviation, however, is Amsterdam Holendrecht, with a forecast that tops the actual ridership numbers by more than 2300. The most accurate results are for Purmerend Wijdevenne in 2009, with a deviation of 17, and for Helmond Brandevoort with a forecast only 92 passengers less than actually generated in 2010.

Table 5.3: Forecasts from the Category 4, Category 5, and Category 6 reference classes

Station	N2009	N2010	Cat4	Cat5	Cat6	Deviation '09	Deviation '10	Ratio '09	Ratio '10
Almere Oostvaarders	3439	3514		2713		726	801	-726	-801
Den Haag Ypenburg	1317	1379		2345		1028	966	1028	966
Arnhem Zuid	1945	2109		2439		494	330	494	330
Voorst-Empe	288	316			564	276	247	276	247
Twello	1330	1460	1173			157	287	-157	-287
Helmond Brandevoort	833	971		902		69	69	69	-69
Gaanderen	339	339	97			242	242	-242	-242
Apeldoorn Osseveld	773	967		2107		1334	1140	1334	1140
Apeldoorn De Maten	636	619		1736		1100	1117	1100	1117
Amersfoort Vathorst	1633	2364		1295		338	1069	-338	-1069
Utrecht Zuilen	1397	1320		2759		1362	1438	1362	1438
Tiel Passewaaij	1230	1152		-72		1302	1224	-1302	-1224
Purmerend Wijdevenne	1578	1269		1595		17	326	17	326
Heerlen De Kissel	419	371		930		511	559	511	559
Groningen Europapark	862	989		1118		256	129	256	129
Eygelshoven Markt	149	285		574		425	289	425	289
Amsterdam Holendrecht	3111	3024		5435		2324	2411	2324	2411
AVERAGE						**704**	**744**	**378**	**309**
CATEGORY 4						**199**	**264**	**-199**	**-264**
CATEGORY 5						**806**	**848**	**468**	**396**
CATEGORY 6						**276**	**247**	**276**	**247**

5.5 Station type Randstad combination division

In this section, the results from the cross-division of the divisions made in the previous two sections. This means that reference classes are composed of homogeneous cases with same values for the variables RANDSTAD, CATEGORY_5, and CATEGORY_6. This cross-division leads to six separate reference classes, of which four are presented in the following subsections. The other two, Randstad Category 4 and Randstad Category 6, are not discussed because the group of 17 stations used to validate the results does not include stations that match these reference classes.

5.5.1 Non Randstad Category 4 reference class

The check for multicollinearity within the Non Randstad Category 4 reference class resulted in the exclusion of eight variables (see appendix 8). Furthermore, the variable URBANIZATION_1 was left out because all stations in this reference class score 0 on it. The variable consequently does not diversify. The first two runs resulted in the removal of the outliers Duiven en Heerhugowaard. The third run was successful and led to the following function:

Y = -890.6 + 152.3 * FREQUENCY_oS + 57.5 * BTM_TOTAL + 212.5 * IC_60 -503.5 * IC_30 + 682.6 * TYPE_SCORE + 0.118 * POP_TOTAL -0.016 * JOBS_MUN

With n = 92
df = 84
R^2 = 0.784
Adj. R^2 = 0.767

The model contains seven variables, of which POP_TOTAL has the most influence on the dependent variable. All variables are significant at, at least, the 95 percent level (see table A8.7 in appendix 8). The model is significant at the 0.000 percent level. The prerequisites regarding linearity and homoscedasticity are met.

5.5.2 Non Randstad Category 5 reference class

Six variables are excluded due to multicollinearity. There are no cases with a standardized residual over 4 within this reference class (see table A9.4). The highest standardized residual belongs to the station Tilburg

Universiteit, formerly known as Tilburg West (see appendix 9). Nevertheless, the model is accepted and the coefficients from table A9.5 lead to the following function:

Y = -1519.1 + 242.6 * FREQUENCY_oS + 2.38 * CROSSINGS + 89.6 * IC_60

With $n = 45$
df = 41
$R^2 = 0.615$
Adj. $R^2 = 0.587$

The model consists of three variables which are all within the 95 percent level. The model is significant at the 0.000 percent level. The requirements for multiple regression have been fulfilled.

The high residual of Tilburg Universiteit station is not remarkable. It is unique within the reference class in the sense that it directly serves the university. Due to the relatively low variance accounted for, it was chosen to add a dichotomous variable for the university with the goal of improving the model. After the addition of the new variable, the following model was calibrated:

Y = -5231.9 + 276.7 * FREQUENCY_oS + 3796.1 * UNIVERSITY + 1.756 * CROSSINGS + 0.335 * INCOME_AV

With $n = 45$
df = 40
$R^2 = 0.759$
Adj. $R^2 = 0.735$

Here too, the model is fully significant, as are the variables at a 95 percent level (see appendix 9). The prerequisites regarding linearity and homoscedasticity were met. The addition of the variable UNIVERSITY has indeed improved the model's quality. IC_60 did not return in the new model. Instead, INCOME_AV is included.

5.5.3 Non Randstad Category 6 reference class

The check for multicollinearity revealed that seven variables needed to be excluded (see table A10.1 in appendix 10). The variables TYPE_SCORE and URBANIZATION_1 were left out because all cases share the same values there. The first MR run led to the removal of Boxtel station in order to achieve improvement of the model's accuracy. The second run revealed no outliers. The resulting function is shown below:

Y = -989.8 + 0.125 * POP_TOTAL + 250.8 * REQUENCY_oS + 59.6 * BTM_TOTAL -105.2 * STOP_30 + 17.0 * TOT_60

With $n = 74$
df = 68
$R^2 = 0.823$
Adj. $R^2 = 0.810$

The model includes five variables, of which POP_TOTAL is the most influential. All variables are significant at, at least, the 95 percent level (see table A10.4 in appendix 10). The model is significant at the 0.000 percent level.

5.5.4 Randstad Category 4 reference class

The correlation matrix revealed that nine variables were to be excluded from the analysis due to multicollinearity (see appendix 11). The first run did not result in any outliers. The function is as follows:

Y = -6481.6 + 816.8 * FREQUENCY_oS + 3658.4 * BTM_BINAIR

With $n = 42$
df = 39
$R^2 = 0.748$
Adj. $R^2 = 0.735$

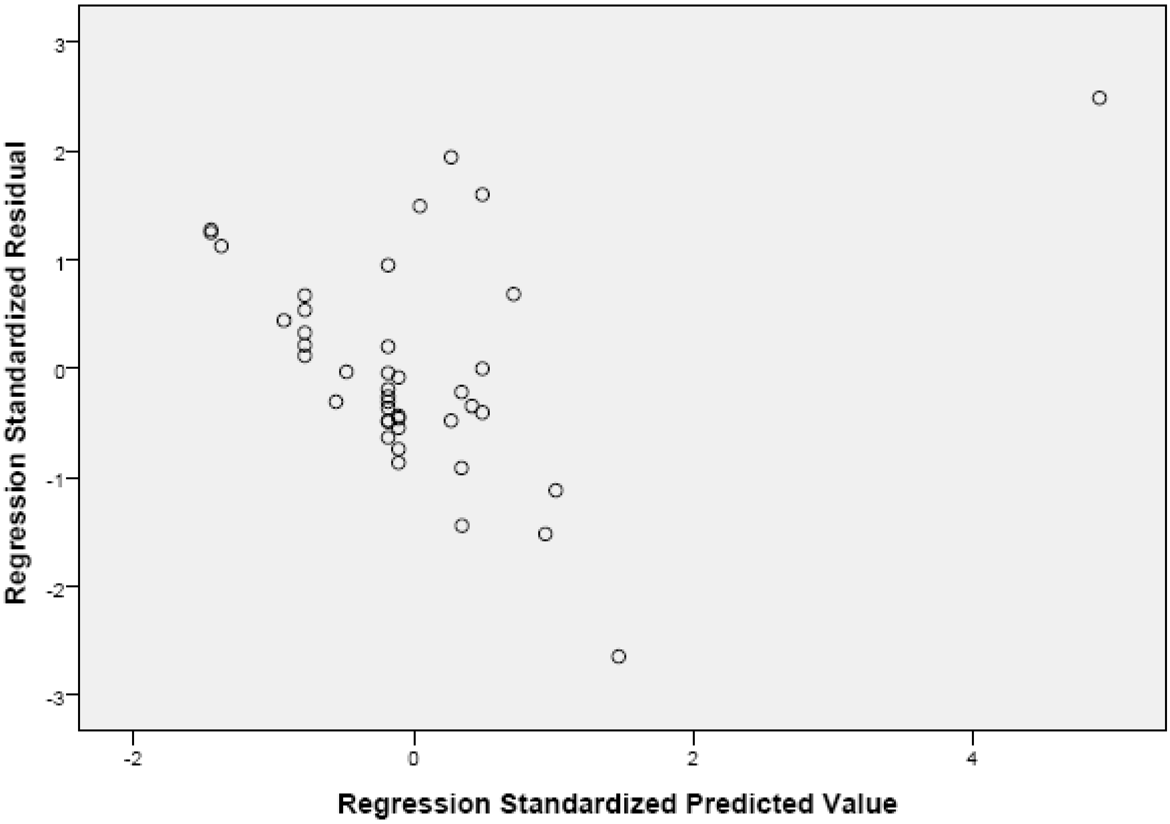
Because the model only includes two variables and POP_TOTAL is not among those, it is attempted to improve, at least theoretically, the model by adding the latter variable through the Stepwise option. This does, nevertheless, not lead to a significant increase in the model's explanatory and forecasting capabilities. The variable is far from significant and only leads to an increase in the model's $R^2$of 0.1 percent (see table 5.4).

Table 5.4: Model summary when POP_TOTAL is included

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,865[a]	,749	,729	3283,998

The initial model is thus used. The prerequisite for homoscedasticity is not unquestionably met though. As depicted in figure 5.2, the pattern is not necessarily random. Instead, one could identify a somewhat linear diagonal. This, however, disregards the outlier in the top right. Still, the model is significant.

Figure 5.2: Scatterplot of the distribution of the standardized residuals



5.5.5 Forecasts for the Randstad Category cross-division

This subsection presents the forecasts from the cross-division between the Randstad/Non Randstad and the categorical division. The forecasts are shown by table 5.5, which contains a column for each of the four reference classes. The average deviation for 2009 for the whole division is 876 and 830 for 2010. The ratios are 728 and 659. The values for the Non Randstad Category 4 reference class, which has two assessment stations among the 17, are remarkably lower. The average deviations are 255 for 2009 and 190 for 2010. The ratio for 2009 is 126 and 61 for 2010. The Non Randstad Category 5 reference class contains ten stations. Of these, Eygelshoven Markt has had a negative forecast. The average deviations are 566 for 2009 and 502 for 2010. They are, consequently, higher than those for the NR Cat 4 reference class. The accompanying ratios are higher as well, with 340 for 2009 and 236 for 2010. As in the categorical division, Voorst-Empe is the only one of 17 stations in its reference class. The predicted ridership number is higher than the averages measured in 2009 and 2010. The fourth reference class, Randstad Category 5, has four stations among the 17 assessment stations. Their forecasts are on average over 2000 passengers over the actual ridership. The forecasts for three of the four stations are equal because the MR function contains only two variables and these stations score the same values for those variables.

The most accurate forecasts are for Purmerend Wijdevenne, with a plus of five for 2009, and for Gaanderen, with a minus of 129 for 2010.

Table 5.5: Forecasts from the Non Randstad Category 4, 5, and 6 reference classes and the Randstag Category 5 reference class

Station	N2009	N2010	NR Cat4	NR Cat5*	NR Cat6	R Cat5	Dev. '09	Dev. '10	Ratio '09	Ratio '10
Almere Oostvaarders	3439	3514				6570	3131	3056	3131	3056
Den Haag Ypenburg	1317	1379				3711	2394	2332	2394	2332
Arnhem Zuid	1945	2109		2587			642	478	642	478
Voorst-Empe	288	316			567		279	251	279	251
Twello	1330	1460	1711				381	252	381	252
Helmond Brandevoort	833	971		544			289	427	-289	-427
Gaanderen	339	339	211				128	129	-128	-129
Apeldoorn Osseveld	773	967		1514			741	547	741	547
Apeldoorn De Maten	636	619		1565			929	946	929	946
Amersfoort Vathorst	1633	2364		2695			1062	331	1062	331
Utrecht Zuilen	1397	1320				3711	2314	2391	2314	2391
Tiel Passewaaij	1230	1152		682			548	469	-548	-469
Purmerend Wijdevenne	1578	1269		1583			5	314	5	314
Heerlen De Kissel	419	371		1069			650	697	650	697
Groningen Europapark	862	989		1363			501	374	501	374
Eygelshoven Markt	149	285		-147			296	432	-296	-432
Amsterdam Holendrecht	3111	3024				3711	600	687	600	687
AVERAGE							**876**	**830**	**728**	**659**
NON RANDSTAD CAT 4							**255**	**190**	**126**	**61**
NON RANDSTAD CAT 5							**566**	**502**	**340**	**236**
NON RANDSTAD CAT 6							**279**	**251**	**279**	**251**
RANDSTAD CAT 5							**2110**	**2117**	**2110**	**2117**

* The numbers in this column are derived from the function given after the inclusion of the university variable (see subsection 5.5.2).

5.6 Results compared

In this section, the results from the various MR functions are compared. This section looks at both the reference classes and the divisions of which they are a result. For both, the indicators for the models' quality are summarized in table 5.6. It contains the three divisions that have been made in order to compose the reference classes. The division averages can be compared to the results of the All Stations reference class to see to which extent creating reference classes can help to increase the accuracy of the forecasts. The average deviation for the All Stations reference class for 2009 is 842. For 2010, it is 799. The decrease shows that it does take newly opened stations a few years to grow to their full potential and that, on average, the stations' ridership numbers are growing towards to forecast values. The average deviations for the Category division are lower, with 704 for 2009 and 744 for 2010. The increase, while a decrease could be expected, can be ascribed to the fluctuations of the generated ridership numbers in the stations' first years of existence. The division averages for the Randstad Category cross-division are higher, with 876 and 830 for the years considered. A closer look at the cause shows that it can be due to the significant over-expectations for the Randstad Category 5 stations. These four stations also cause the deviation averages for the Randstad division to be relatively high, since they belong to the Randstad reference class. This reference class resulted in forecasts topping generated numbers by, on average, more than 2500 for both years considered.

The ratios, which give insight in the bias of the models, for the various divisions are all higher than the average ratios for the All Stations reference class. The ratios of the Category division come closest to the ones of the All Stations reference class, followed by those of the Randstad division. The ratios for the Randstad Category cross-division are the furthest away. Again, these numbers, 728 for 2009 and 659 for 2010, are skewed by the over-optimistic forecasts for the Randstad Category 4 stations. The ratios are supposed to show a plus of a few hundred because the stations are expected to generate higher ridership volumes over the coming few years. For the smallest stations, this growth may be limited to several dozens, while the largest stations can show an increase of over 1500 within the first five years (Projectteam toepassing norm, 2009). The value of these numbers will be discussed in chapter 6.

Table 5.6: Forecast results for the divisions and reference classes

		Function		Average deviation		Average ratio		Negative
Division	Reference class	*n*	R^2	2009	2010	2009	2010	forecasts
None	All Stations	291	0.770	842	799	313	245	3
Randstad	Randstad	79	0.736	2565	2571	2565	2571	0
	Non Randstad*	214	0.720	625	647	90	8	2
	Division average			1325	1378	402	333	3
Category	Category 4	123	0.781	199	264	-199	-264	0
	Category 5	84	0.722	806	848	468	396	1
	Category 6	82	0.786	276	247	276	247	0
	Division average			704	744	378	309	1
Randstad Category cross	Non Randstad Cat 4	92	0.784	255	190	126	61	0
	Non Randstad Cat 5	45	0.759	566	502	340	236	1
	Non Randstad Cat 6	74	0.823	279	251	279	251	0
	Randstad Cat 5	42	0.748	2110	2117	2110	2117	0
	Division average			876	830	728	659	1

*Averages excluding Groningen Europapark station (see subsection 5.3.3)

Compared to the reference classes, All Stations does not have the lowest ratios. The Non Randstad reference class does, if the case of Groningen Europapark is disregarded. Otherwise, the Non Randstad Category 4 reference class has the lowest ratios for both 2009 and 2010. This reference class is also among four reference classes with the lowest average deviations. The others are Category 4, Category 6, and Non Randstad Category 6. The reference classes with the highest average deviations and ratios are Randstad and Randstad Category 5.

The reference class with the highest variance accounted for is Non Randstad Category 6, with an R^2 of 0.823. This means the significant variables combined with the intercept account for 82.3 percent of the variance in the ridership numbers for category 6 stations located outside of the Randstad. The other reference classes with variance accounted for over 78 percent are Category 4, Category 6, and Non Randstad Category 6. The lowest R^2 is for the Non Randstad reference class. The corresponding MR function explains 72.0 percent of the variance in Y. A high R^2 does not necessarily lead to accurate forecasts, because the new case at hand might have specific characteristics that affect the actual ridership numbers that are not taken into account by the MR function because this characteristic was lacking among existing stations. Nevertheless, a high R^2 might be an indicator for, on average, higher accuracy. A higher *n* is neither a guarantee for high accuracy. The Non Randstad reference class, with the second highest *n* and close to zero ratios, has the lowest R^2 and relatively high average deviations. On the other hand, a lower *n* is not the way to high accuracy either, judging by the values for the Randstad Category 5 reference class.

Another measure for the quality of the forecasts per reference class is the number of cases that received a negative (as in below zero) forecast. As table 5.6 shows, none of the divisions go without negative predictions. The Category division and the Randstad Category cross-division both have a single negative forecast. The former has a forecast of -72 for Tiel Passewaaij (see table 5.3) while the latter has a forecast of -147 for Eygelshoven Markt (see table 5.5). The Non Randstad reference class resulted in three negative forecasts, of which -3738 for Groningen Europapark was the most extreme. The All Stations reference class also led to three negative forecasts. Mathematically, negative forecasts are not impossible, as they are simply the outcome of the functions provided. Practically, however, negative ridership numbers are impossible. The application of prediction intervals to the forecasts might be a solution to avoid negative forecasts. Prediction intervals can be acquired through SPSS, together with the MR outputs. The application of prediction intervals is discussed in chapter 6.

5.7 Conclusion

The stations Helmond Brandevoort, Tiel Passewaaij, and Eygelshoven Markt are apparently hard to forecast accurately. All received negative forecasts in two separate reference classes. In one instance, Groningen Europapark also had a forecast below zero. The category 5 stations of large cities are hard to accurately forecast anyhow. Aside from the Randstad Category 5 reference class, the reference classes that originate from the Randstad Category cross-division seem to provide the most accurate results. The results have to be weighed carefully though, considering that it can be expected that the 17 assessment stations will generate higher ridership volumes over the next several years. This may vary from several dozens for the smallest stations to several hundreds or even 1500 for the larger category 5 stations. It has become clear that a high variance accounted for is no guarantee for accurate forecasts, since all R^2 values are in the 0.7s, except for one which is over 0.8. The *n* number of cases in the reference class does also seem not to be directly related to the accuracy of the forecasts. The forecasts as presented in this chapter will be discussed in chapter 6.

In the Randstad Category cross-division, the station Eygelshoven Markt gets a negative forecast. This is partly due to the intercept of minus 5231.9. In addition, the frequency of service at the station is only two, as opposed to the more common frequency of four for most stations. The reference class' function further includes the variable INCOME_AV. Though it is possible that a higher household income is associated with higher ridership at the station level because it is related to jobs and thus to mobility, it can be questioned whether it is a good driver for ridership in this function. Firstly, it was only included after the creation of the university variable. Secondly, this is the only one among the reference classes' functions that contains this variable. A detailed discussion of the results is provided in the following chapter.

6 Analysis and discussion of the results

6.1 Introduction

In chapter 5, the results of the multiple regression functions for the various reference classes are presented and compared. This chapter will build on the findings from chapter 5 by discussing these results and by paying in depth attention to the most remarkable outcomes. Firstly, section 6.2 discusses the role of the variables in included in the functions, per reference class, on the ridership forecasts of the assessment stations in the respective classes. Secondly, the negative forecasts are analyzed, with the purpose of finding the sources for such inaccuracies. Section 6.4 subsequently analyzes and discusses the highly over-optimistic forecasts for the same reasons. In section 6.5, the use of the reference classes is discussed. Recommendations for modifications that may lead to more accurate forecasts are summarized in section 6.6 before this chapter is concluded in section 6.7.

Table 6.1: The variables included in the MR functions per reference class

Variable	+	-	±	All Stations	Randstad	Non Randstad	Category 4	Category 5	Category 6	Non Randstad Cat 4	Non Randstad Cat 5	Non Randstad Cat 6	Randstad Cat 5	COUNT
POP_TOTAL	x					x	x		x	x		x		**5**
ONEP_HH		x		x										**1**
CROSSINGS	x			x							x			**2**
BUILTUP_P	x			x				x						**2**
TOT_60	x						x					x		**2**
IC_60	x			x		x		x		x				**4**
TOT_30		x		x		x								**2**
IC_30			x					x^+		x^-				**2**
STOP_30		x							x			x		**2**
FREQUENCY_oS	x			x	x	x	x		x	x	x	x	x	**9**
BTM_BINAIR	x				x								x	**2**
BTM_TOTAL	x			x		x	x	x	x	x		x		**7**
TYPE_SCORE	x			x		x	x			x				**4**
TRANSFER	x			x			x							**2**
JOBS_MUN			x		x^+					x^-				**2**
INCOME_AV	x										x			**1**
URBANIZATION_1		x				x								**1**
URBANIZATION_3	x								x					**1**
URBANIZATION_4		x		x										**1**
CATEGORY_5		x		x	x	x								**3**
UNIVERSITY	x										x			**1**
COUNT	**13**	**6**	**2**	**11**	**4**	**8**	**6**	**4**	**5**	**7**	**4**	**5**	**2**	

6.2 Variables' summarized

The ten multiple regression functions presented in chapter 5 are all in different compositions containing different variables. The variables with at least one appearance in the MR functions are included in table 6.1. It shows that, in total, 21 variables made it into the functions. Of these, 13 have a positive influence on ridership, six have a negative influence, and two have a positive and a negative influence both in one occasion. Of the 21 variables, six make only one appearance and nine make only two appearances. The most prevalent variable is FREQUENCY_oS. The only function it is not included in is that of the Category 5 reference class. The second and third most prevalent variables are BTM_TOTAL and POP_TOTAL. The frequent occurrence of these variables is logical and argumentative. Firstly, the more people within a station's sphere of influence, the more potential a stations has to draw a large crowd. Secondly, the higher the frequency of service by the underlying public transit network of busses, trams, and subways, the better the opportunities to travel to and from the station. Thirdly, a high frequency of service makes travelling by train more attractive due to reduced average waiting time and better opportunities for intermodal transfers. A combination of these three variables is found in Non Randstad, Category 4, Category 6, Non Randstad Category 4, and Non Randstad Category 6.

In two cases, the variable CROSSINGS is included. This variable was originally included as a measure of accessibility of the station. The correlation matrices showed, however, that CROSSINGS is highly related with POP_TOTAL. In other words, because of the high correlation between the number of intersections and the number of residents it is unlikely that the two will be included in a single model together. Due to the high correlation, both basically account for almost the same variance in ridership. It can therefore be stated that in the two reference classes that have a function containing the variable CROSSINGS, it serves as a substitute for the population variable rather than as a measure for accessibility.

The variables TYPE_SCORE and IC_60 both make four appearances. The former is a dummy for the service of trains other than stop trains. These are usually intercity trains that also stop at non intercity train stations as part of the regular schedule. The stopping of intercity trains has a positive effect on ridership because it grants passengers quicker access to main intercity stations. These are most frequently the destinations of people accessing the train at local stations. The latter variable, IC_60, is a direct measure for the accessibility of intercity train stations from the stop train stations. A high number of intercity stations accessible within an hour will lead to higher ridership for the origin station because it gives people more destinations within a reasonable amount of time. More destinations means that the train becomes an alternative for the automobile for more people. The same argumentation applies to the variable TOT_60, although this number also contains other stop train stations.

Two of the other stations accessibility variables, being TOT_30 and STOP_30, which are highly correlated, have a negative effect on ridership. This makes sense, because when a station is located in between a high number of other stations, it will have a smaller hinterland than a station without many stations in its proximity. The stations are thus each other's competitors. The IC_30 variable has a negative effect in the function for the Non Randstad Category 4 reference class, but that seems to be more of a mathematical correction for the IC_60 variable, which is also included in that function. In the Category 5 reference class' function, IC_30 has a positive influence. This might be explained by the relative position category 5 stations have with respect to intercity stations. Category 5 stations are often located in dense, urban areas where they provide access to the larger intercity stations within a relatively short distance.

The results presented in chapter 5 show that the forecasts for category 5 stations are not always accurate. Three category 5 stations received negative forecasts in multiple occasions and the ridership numbers of some of the stations in large cities were grossly overestimated. The summary of the variables in table 6.1 shows that category 5 stations generally generate lower ridership numbers than other stations do. The variable CATEGORY_5 has a negative influence in three functions. These three functions are the only ones the variable can be included into because all other reference classes are based on the station categories, which means that this variable is not diversifying in these classes. To find the causes for the inaccuracies in the forecasts, they need be analyzed in depth.

6.3 Analysis of the negative forecasts

In the All Stations reference class, Helmond Brandevoort, Tiel Passewaaij, and Eygelshoven Markt all have negative forecasts. This can, at least partially, be attributed to the gradient accompanying the category 5 variable. The value of the gradient is negative 1259.4. The height of the gradient is an outcome of the multiple regression analysis in SPSS and is averaged out to capture the relatively lower scores for category 5 stations, including those in large cities. These stations, such as Arnhem Zuid, Den Haag Ypenburg, and Amsterdam Holendrecht, happen to have gotten overestimated forecasts of over a thousand. This leads to the assumption that category 5 stations perform differently in large cities than is smaller cities. The division based on the station categories is thus, as such, insufficient to capture the effects that lead to the relatively lower ridership numbers for category 5 stations. Subdividing of the Category 5 reference class based on the Randstad characteristic is neither sufficient, since large cities such as Amersfoort, Arnhem, and Groningen are located outside of the Randstad.

The negative forecast for Tiel Passewaaij in the Category 5 reference class seems to have a different cause. The variable IC_30 is included in its function, but there are no intercity stations within 30 minutes of Tiel Passewaaij. The closest stations are 's-Hertogenbosch and Utrecht Centraal, but these are both just over 30 minutes. Oss and Tilburg are both a bit further away. Anyhow, because Tiel Passewaaij scores zero for this variable, the variables in the function altogether cannot make up for the negative intercept and the result is thus a negative forecast. The IC_30 variable is also responsible for the underestimated forecast for Amersfoort Vathorst. The 2010 deviation is over 1000 off, which would have been less in case the station had a score for IC_30 higher than one. The other cause for the underestimation is the low built up percentage of the surroundings of the station. The area has been under development over the past years, so there is a good chance that the 2008 dataset is missing certain urban expansions that have positively affected the ridership numbers for the station.

The negative forecasts of the Non Randstad reference class can most likely be attributed to the dummy variables CATEGORY_5 and URBANIZATION_1. In the case of Groningen Europapark, the latter is especially relevant because the station scores one for this variable. This means that the variable's gradient (-4194.2) plays a large part in the station's forecast of minus 3738. Given the fact that the Non Randstad reference class contains only one other station with the same characteristic, it is probably better to ignore this variable when forecasting. There is no guarantee that the characteristic of a high degree of urbanization would lead to lower ridership numbers. To the contrary, literature suggests that a high degree of urbanization affects rail ridership positively. The reason the variable was kept in the function in the first place has been the possibility that it could account for certain unmeasured characteristics. It can be concluded here however, that in order to avoid coincidence and randomness from affecting forecasts, variables of these kinds should be avoided in the forecasting of rail ridership.

Aside from Groningen Europapark, two other stations in the Non Randstad reference class have gotten negative forecasts. These are Helmond Brandevoort and Eygelshoven Markt. For the latter, the effect of the dummy variable CATEGORY_5 is thought to play a part. The accompanying gradient of minus 1040 can be a lot to catch up to for a small station, especially when the function also has a negative intercept. The negative score for Helmond Brandevoort has a different, more data related cause. As of 2011, the district Helmond Brandevoort has over 8000 inhabitants, all living relatively close to the station. The dataset used here, however, includes 2006 census data. After the application of the 2000 meter network distance buffer, it was found that the stations sphere of influence contained only 2987 residents. Given the recent developments, the low forecast should therefore be put in perspective of the difference in population data for 2006 and 2011. Adjustment of the population data will lead to a higher, more accurate forecast because the variable POP_TOTAL is included in the MR function for this reference class.

Another option to avoid negative forecasts is to express them in a bandwidth, rather than a single number. In statistics, such a bandwidth is known as a prediction interval. This is an estimate of the interval, based on the existing observations such as in the reference classes, in which a proposed or future case's forecast will fall. A prediction interval always has a certainty to it, expressed as a percentage. This could be, hypothetically, that the ridership value for station X, will fall, with a 70 percent probability, between 1200 and 1800 daily passengers. One benefit of using a prediction interval for a forecast is that the chances the actual, generated value will fall within the interval are much higher than a single a single value forecast being the same as the generated value. In other words, it provides more certainty. In the case of a negative forecast, the top end of the accompanying prediction interval might have a good chance of being positive. This will then give a better indication of the ridership to be expected than the single value negative forecast. For these reasons, it can be advisable to express forecasts as an interval. A disadvantage of using a prediction interval is that it can become very wide when a high probability is pursued. The individual prediction interval may grow into the thousands for a 90 percent probability. For a station with an expected ridership volume of only several hundreds, such a wide interval can be anybody's guess. Another disadvantage is that in case of multiple regression, it requires statistical software such as SPSS to calculate a forecast's interval. This exclusivity makes it unapt to apply to a GIS based method.

6.4 The highly over-optimistic forecasts

Determining which of the forecasts for the 17 assessment stations are highly over-optimistic is delicate. For one, it should be taken into account that the ridership volumes, generated in 2009 and 2010, are still growing before having reached their potential. In the example of Amersfoort Vathorst in the Randstad Category cross-division, the forecast can be considered overly optimistic when set off against the actual ridership number of 2009 (2695 versus 1633), while it can be considered fairly accurate when compared to the 2010 ridership number of 2364. What also needs to be taken into account is the size of the station. When the forecast for Gaanderen exceeds the actual ridership by 600-800 for example, this can be considered a large overestimation, while the same difference for a station like Amsterdam Holendrecht is relatively smaller.

With these considerations in mind, one can value the forecasts and judge whether they are overly optimistic or not. Doing so reveals that Utrecht Zuilen and Den Haag Ypenburg appear over-optimistically forecasted in all four occasions, with forecasts exceeding actual ridership from around 1000 up to almost 3000. Other stations with multiple highly over-optimistic forecasts include, but are not limited to, Almere Oostvaarders, Arnhem Zuid, Apeldoorn De Maten, and Amsterdam Holendrecht. This shows that the variables used in the multiple regression analyses were not capable of accounting for an important part of the variance in ridership numbers for the category 5 stations of large cities. Theoretically, the Randstad Category 5 reference class should have been able to capture the unexplained part because of the internal homogeneity of the reference class. It turned out however, that it could not, judging by the grossly over-estimated forecasts in this reference class. The function for this reference class contains only two variables, which shows that the variables were either unable to account for the variance or that there is a high correlation between the variables in the function and the insignificant variables.

The multiple inaccurate results for most of the category 5 stations indicate that there are certain influences on these stations' ridership that are not adequately measured by the variables used. A discussion of these influences is, of course, speculative, but the results presented in chapter 5 do provide indications of lacunas in the data and provide indications for possible adjustments.

The functions for the Randstad reference class and for the Randstad Category 5 reference class, for starters, both include the variable BTM_BINAIR. This indicates that the BTM_TOTAL, the variable for the amount of busses, trams, and metros stopping at the station, is less apt to explain the variance in ridership than the variable containing the characteristic of BTM stopping or not at the station. BTM_TOTAL's incapability to do so can have two causes. First, the total of BTMs at certain stations in the reference classes might be so high that it stops adding to the variance. In other words, when the number of BTM's stopping at a station goes beyond a certain value, it stops adding to the ridership in a linear way. When there is a significant number of stations with BTM values beyond that threshold value within the reference class, the variable's effect on ridership can become impossible to depict. Transforming the variable into an ordinal scale could help to overcome this.

A second cause can be found in the contradiction between feedering and competing public transit. Where feedering public transit draws passengers to a train station, competing public transit can lure them away by providing an alternative to travelling by train. Competing public transit is especially relevant for highly urbanized areas with dense networks of underlying public transit. In these cases, bus routes, tram lines, and subways parallel to a train connection can be worthy alternatives to the train. As such, even a low number of BTMs can have a negative effect for certain stations. Because it has been impossible to diversify between feedering and competing BTMs, this effect of underlying public transit could not be captured in the multiple regression analyses. This effect might contribute to the overestimated forecasts for the stations Utrecht Zuilen and Den Haag Ypenburg, however, since both stations are located within urban areas that are served by multiple modes of high-frequency public transit.

Another aspect in the overestimation of ridership numbers for category 5 stations might be the determination of their service areas. In cases of category 5 stations that are supplementary to intercity stations, the way in which Utrecht Zuilen is supplementary to Utrecht Centraal and Den Haag Ypenburg to Den Haag Centraal and Den Haag HS, the service areas of the category 5 stations might have been overstated. It was assumed that people would choose the closest train station when the service areas where calculated. Obviously, this is not always the case, since the direction of the trip, access mode, custom, and the station's frequency of service, among others, might lead to other considerations. In cases of category 5 stations near other stop train stations, such as the case of Nijverdal and Nijverdal West, these considerations will have a limited effect. In cases of category 5 stations supplementary to an intercity station, they might have more effect, though. This could result in service areas that are smaller than they are in this study. ArcGIS, however, does not provide the means to simultaneously execute a Network Analyst service area calculation in which one input station is weighed more heavily than another, which would result in service areas of different magnitudes.

6.5 The use of reference classes

Regarding the forecasts for category 5 stations, based on the findings from the previous sections 6.3 and 6.4, it can be stated that they are frequently inaccurate with the application of the reference classes as they were. The reference classes that do not regard category 5 stations have, nevertheless, resulted in fairly accurate forecasts, especially when compared to the forecasts in the All Stations reference class. The average deviations for the reference classes that result from divisions are lower, but seem to remain positive enough to leave some space for the ridership numbers that are growing within the first few years of the stations' existence. As such, the application of reference classes is shown to have a positive effect on the accuracy of the ridership forecasts.

The most accurate forecasts for non Randstad stations seem to be gotten from the Randstad Category cross-division. The Non Randstad Category 4 reference class and the Non Randstad category 6 reference class seem especially apt for forecasting. Not only do they have high variance accounted for, but also do they yield to forecasts that are, on average, a few hundred above the generated ridership in 2009 and 2010. Because the gap between forecasts and actual numbers is getting smaller, it can be stated that ridership numbers for these stations are growing towards the forecast values. The reference classes resulting from the Randstad Category cross-division are relatively small and have high variance accounted for. This means that the variables can account well for the variance in ridership due to the homogeneity of the reference class. Further subdivision of the reference classes is nevertheless not considered, because that would negatively affect the applicability of the regression functions for forecasting. A lower amount of cases increases the chances of random errors in the regression functions. The Randstad Category cross-division is, therefore, well balanced between population size and homogeneity of the reference classes on the one hand and statistical reliability on the other.

Considering the less accurate forecasts for category 5 stations, it is proposed here to pursue subdivision of these stations based on different criteria. Apparently, diversification based on a station's position within or outside of the Randstad does not add to the accuracy of the forecasts, especially not for stations located in large cities as supplementary to an intercity station. As a consequence, two criteria for further partition of the category 5 stations seem more adequate. First, a subdivision based on the size of the stations' cities can be considered. Because it turns out that the forecasts for large city suburban stations are especially inaccurate, it would make sense to apply this criterion. The problem with this criterion is, however, that it is unclear what constitutes a 'large' city. Put differently, it is not known which threshold value for a city's population number should be applied for this subdivision. Applying this criterion would therefore require further research. The second possibility for a partition of the category 5 stations class is a division based on the stations' relative position to an intercity station. This would results in a group of stations which are supplementary to an intercity station and a group of category 5 stations which exist concurrently with one or more other stop train stations. Examples of the latter include Barneveld Noord, Delfzijl West, Emmen Zuid and Sneek Noord. The downside of this division is that it would results in a relatively small number of stations in the latter reference class. Furthermore, this would also require further research, because it is currently unclear how a position relative to an intercity station should be measured and expressed.

As discussed in chapter 2, the advantage of using reference classes for forecasting is that it rules out human bias and deliberate manipulation of outcomes to a large extent. Although deliberate manipulation is never fully ruled out, the use of reference classes, as shown by chapter 5, does seem to lead to less overly optimistic forecasts. The ratios are on average only a few hundred over the ridership numbers realized thus far and there is also one that is under the generated ridership. This shows that the model as such is not biased. Manipulation of the results is still possible through the manipulation of data. As shown by the case of Helmond Brandevoort, obsolete, or otherwise incorrect data can have a significant effect on the forecast. In the case of Helmond Brandevoort, this has not been deliberate, but rather it is a consequence of using 2006 data for a forecast subsequently compared to the generated ridership of 2010. What this example shows is that accurate, up-to-date data are required to gain accurate forecasts. The principle of garbage in, garbage out is thus especially true for ridership forecasts.

6.6 Recommendations put to practice

The discussion in the sections 6.2 through 6.5 brings forward a number of possibilities for improvement and recommendations to increase the accuracy of ridership forecasts. For one, it has been shown that the use of certain variables, especially dummy variables, can be serviceable to account for variance in ridership for the existing stations, but is less apt to use in forecasting. They can results in highly over-optimistic or negative forecasts. Since both are unintended, it is recommended to not base forecasts on variables that can have certain effects. Secondly, it has been shown that ridership numbers for category 5 stations in larger cities are difficult to accurately forecast with the variables used here. Either a different division for reference classes should be applied, or different variables, such as separate variables for feedering and competing public transit, should be used. Both, but the latter especially, require more research. The use of reference classes has indeed been shown to be helpful in pursuing higher accuracy in ridership forecasts. Thirdly, the use of individual prediction intervals has been suggested to achieve a higher degree of certainty for the individual forecasts. In this section, it is attempted to implement the previously discussed and proposed modifications.

The proposed modifications that do not depend on further research have been applied to see their effects. This means that the category 5 stations of cities over 110.000 inhabitants, such as Den Haag Ypenburg and Groningen Europapark, are not included. The results are summarized in table 6.2. The table includes a column with the forecasts resulting from the modifications and columns to show the effect of the individual prediction intervals. The forecast for Voorst-Empe is identical to the forecast it received from the Non Randstad Category 6 reference class because the recommendations do not affect it (see chapter 5, table 5.5). The modifications for Twello and Gaanderen are limited to an adjusted regression function, because some of the variables that were included are replaced at the gain of other for the purpose of forecasting, as discussed earlier. The forecasts for the remaining five stations are based on a new reference class, namely one based on all category 5 stations located in cities with less than 110.000 inhabitants. This inevitably means that a new multiple regression function is drawn up. For the purpose of comparison, the forecasts from the Randstad Category cross-division are also included.

Table 6.2: Results after implementation of proposed modifications

	Ridership		NR Cat	Interval (20%)			Interval (40%)		
Railway station	**2010**	**Forecast**	**forecast**	**Low**		**High**	**Low**		**High**
Voorst-Empe	316	567	567	449	-	687	321	-	815
Twello	1460	1882	1711	1644	-	2120	1389	-	2376
Helmond Brandevoort	971	484	544	250	-	718	0	-	969
Gaanderen	339	285	211	44	-	527	0	-	785
Tiel Passewaay	1152	871	682	643	-	1099	397	-	1345
Purmerend Weidevenne	1269	2033	1583	1804	-	2263	1557	-	2509
Heerlen De Kissel	371	935	1069	705	-	1164	460	-	1410
Eygelshoven Markt	285	340	-147	108	-	574	0	-	824

Table 6.2 shows that the modifications have resulted in the absence of negative results. Four forecasts are higher than in the Randstad Category cross-division, three are lower, while the forecast for Voorst-Empe remains unchanged. Four forecasts have become closer to the reality of 2010, while three are further off. It should be emphasized here that this provides no more than a rough indication of the accuracy, since it may take stations several years to grow to their full potential.

The last columns of table 6.2 show the prediction intervals for the individual forecasts. As already discussed in section 6.4, a higher probability leads to widening intervals. The table includes the 20 percent and the 40 percent intervals. Dependent on the reference class, the intervals are about 200 to 500 wide with a 20 percent probability, and 500 to 1000 for the 40 percent interval. Because the intervals for the 40 percent probability intervals are already so wide, pursuing a higher probability is unnecessary. The intervals will become so wide that anyone can predict the ridership will fall within it. On the other hand, a probability of only 20 percent appears to be rather low.

6.7 Conclusion

Evaluation of the variables that account for the variance in ridership learns that in addition to the number of residents within the station's service area, frequency of service and frequency of service by busses, trams, and metros are the most important drivers of ridership in The Netherlands. Furthermore, the number of destinations, especially the number of intercity stations within 60 minutes, is important, while the number of stop train stations within 30 minutes can have a negative effect. The latter can be attributed to mutual competition for passengers. Depending on the reference class used, other variables may also account for ridership at the station level. Not all variables seem equally apt for forecasting though. The dummy variables, especially, can have a negative impact on the accuracy of forecasts in cases of unprecedented values for the proposed station or unprecedented or unusual combinations of variables.

Although the Randstad Category cross-division generally leads to fairly accurate forecasts, there are a number of possible modifications in order to achieve higher accuracy. The reference classes' regression models achieve fairly good variance accounted for, with a high in the 80 percent, but this still leaves room for improvement. Adding certain variables will, however, add relatively little to the variance accounted for, since it is already quite high. An important conclusion is that accurately forecasting ridership for the category 5 stations located in large cities are difficult to make. More research will be required to find causes and solutions. Other modifications, such as the exclusion of certain variables for aforementioned reasons, seem to result is higher accuracy. Statements about accuracy in this case need to be made with care though, since the assessment stations have not had the time yet to fully mature and reach their full potential. With more certainty, it can be stated that the model is less susceptible to deliberate manipulation or fraud than conventional methods because there is less room for human adjustments. Only the data acquisition and adjustment leave room for manipulation. The application of accurate data is thus very important to insure accurate forecasts. Furthermore, it is found that the model is not biased. Current results indicate that the ratios of forecasts versus generated ridership numbers will be, after maturing of the stations, close to zero.

7 Conclusions

7.1 Conclusions

Recent years have seen the opening of a large number of new train stations in The Netherlands. In total, 22 new stations have been constructed and become operative between 2004 and 2010. More will be opened in 2012 and many more have been proposed for the subsequent years. In the decision-making process that leads to the construction of a new station, ridership expectations are usually a decisive factor. Passenger flows are important for a station's economic viability because they are a measure for the station's revenues. When demand shortfalls occur, the station's revenues are threatened. In these cases, the construction of the station can be considered to be an unprofitable use of public funds, which could have better been spent elsewhere. Demand shortfalls often occur in The Netherlands, as conventional methods used by commercial forecasters tend to be biased towards over-optimism. For these reasons, pursuing an unbiased, more accurate method is justified. This study was aimed at finding the building blocks for such a method. The study has adopted a number of aspects to be playing a part in an improved method:

1. The principles of reference class forecasting
2. Network based service areas of two kilometer
3. Multiple regression (MR) to identify significant variables
4. MR functions to make forecasts
5. Inclusion of the above aspects in a GIS environment

With regards to the first of these aspects, it was found that conventional methods lead to overly optimistic forecasts for three reasons. First, the use of inaccurate data results in errors in forecasts. This problem is described with the phrase garbage in, garbage out. This problem, however, only accounts for random errors and does only partially explain the tendency towards over-optimism. Inaccurate input data also lead to overly pessimistic forecasts. This has been shown by the case of the station Helmond Brandevoort, which is located in an area of recent urban expansion. The population data of 2006 do, in this case, not fully account for the station's ridership in 2010 because in the mean time, the number of residents within the station's sphere of influence has grown. This results in underestimated ridership numbers. Secondly, a human characteristic known as optimism bias is found to be causing overly optimistic results. This is the human tendency to adopt or be affected by optimism of others. Practically, this can be explained by the following example. When a local decision-maker approaches a forecaster with the assignment to forecast ridership for a new station, then that decision-maker is most likely positive about the stations effects. If not, then the project for the station would have been shut down earlier anyway. His positive attitude subsequently affects the forecaster, who, in turn, unconsciously affects the outcome is a positive way. A third cause for over-optimism is the deliberate manipulation of forecasts, which is known as strategic misrepresentation. This is prevalent when multiple parties compete for the same funds. Because the project that is most beneficial will walk away with the funding, it is in all competing parties' interest to make their project look as positive as possible. High expectations of ridership, as a consequence of manipulation, will help to make it look that way. Another situation in which strategic misrepresentation can emerge is when a commercial party that is doing the forecast can expect to gain more work, and thus profit, out of a positive advice. For the reasons discussed here, it can be concluded that a method aiming at higher reliability should leave little to no room for optimism bias and strategic misrepresentation.

The proposed solution for optimism bias and strategic misrepresentation is known as reference class forecasting (RCF). The principles of RCF include using historical precedent to base forecasts on. This means that a group of similar cases, which together make up the reference class, should form the basis for the forecast. Furthermore, forecasters need to refrain from focusing on the details of the project at hand, such as the resources available, expected scenarios, and possible hiccups. Forecasters should especially avoid creating images of the expected future developments in their mind. If these conditions are met, expectations for the investigated case can be extrapolated from the reference class based on statistics. Reference class forecasting is proven to be unbiased, less susceptible to strategic misrepresentation, and thus yields unbiased forecasts.

Aside from their susceptibility to optimism bias and strategic misrepresentation, conventional methods for ridership forecasting are characterized by the use of crow-flight distance buffers to determine stations' service areas. The use of these buffers is associated with the overestimation of ridership, since they overstate the sphere of influence of a station. It has been shown that actual, network distance buffers generated by GIS lead to smaller calculated service areas for stations. Furthermore, such service areas vary in size, which means the difference between a network distance buffer and a crow-flight buffer varies per station. It is concluded that the use of road network distance buffers is more accurate and can therefore help to achieve more accurate results. The distance value used for the buffers was set at two kilometers. This distance is chosen because it accounted for higher variance in ridership than 500, 1000, 1500, 3500, and 5000 meter buffers, based on the number of residents within the buffers. The exact number of residents within the buffers is calculated by calibrating the number of residents within the four-digit postal code zones onto the areas with a residential land use. This assures an accurate representation of the actual number of people living within the service area of a station. Not just the number of residents is calibrated this way, other variables, such as the number of road intersections, the ratio between residential and commercial land use types, and car ownership are also generated through the network-based service areas.

In total, 30 variables associated with train ridership at the station level were identified based on literature. It was managed to gather sufficient data to quantify 17 of these, resulting in 31 variables used as input for the multiple regression analyses for the various reference classes. From the multiple regression analyses, it was concluded that the number of residents, frequency of service by train, frequency of service underlying modes of transit, and accessibility of destination stations are the most important and most prevalent drivers of ridership. Because of high correlation between certain variables, the same variables did not always appear in the MR functions. The crossings variable and the total population variable strongly correlated, for example, which made it unlikely for both to be included in the same function. It was further shown that the number of train stations within 30 minutes has a negative effect on a station's ridership. This is most likely attributable to the mutual competition for the same passengers. From the All Stations reference class, it was shown that having more than two directions to travel to and being serviced by intercity trains generally adds to ridership of a station. An important conclusion is that category 5 stations, especially those that are subsidiary to intercity stations, generate relatively - as opposed to the number of residents within their service areas - smaller passenger flows than other stations.

All reference classes had MR functions with high R^2 values, which means they account fairly well for the variance in ridership. The division of all stations that led to the reference classes with the highest variance accounted for was the Randstad Category cross-division. This discriminated between stations based on their location within or outside of the Randstad on the one hand, and based on their categorization on the other hand. A high R^2 does not necessarily mean that the corresponding regression function leads to accurate forecasts, however. This has been shown by the Randstad Category 5 reference class. While the regression function accounted for 74.8 percent of the variance of the ridership of the stations within the reference class, the forecasts derived from the MR function were highly off. Since category 5 stations of larger cities outside of the Randstad also turned out to be difficult to accurately forecast, it has been concluded that the variables that were used to try to forecast were not enough to come to highly accurate forecasts. Furthermore, different composition of the reference class for these stations could be considered.

When looked at the method as a whole, it is concluded that it is unbiased, in the sense that the forecasts are both above and below eventually generated ridership volumes. Although the balance for the 17 assessment stations was leaning towards optimism on average, it is expected that this balance will shift towards the middle because the ridership numbers of the new stations will grow over the coming years. It takes a new station approximately five years to grow to a stabilized ridership number, but even then, large changes can occur. It is for these reasons that statements about the accuracy of forecasts need to be made with care and mindfulness. Nevertheless, it can be concluded that the accuracy of the forecasts, despite unbiased results, could become higher. An advantage of the model is that it leaves little room for the deliberate manipulation and misrepresentation of forecasts, because it is the historical precedent of the reference class that is decisive in the forecast, rather than detailed input data and assumptions based on the case at hand. Of course it is possible to tamper with data, but that is no different from other methods.

The principle of reference class forecasting, network-based service areas, and the SPSS originated multiple regression function can all be included into a GIS environment. AcrGIS's Model Builder provides means to construct a model that is able to calculate a forecast for any new station in The Netherlands. However, calculating the network-based service areas is forced to take place outside of the model, because it requires a road database to be converted to file containing M-values. Furthermore, when there are indications that adjusting a reference class may lead to improved forecasts, when updated data becomes available, or when new variables can be applied, it requires a shift to SPSS to recalculate the MR functions used. Subsequently, this can require the construction of new model. That is a major limitation.

What also requires SPSS for their determination, are the individual prediction intervals for station forecasts. In practice, a station's generated ridership number will hardly ever be exactly equal to the predicted value. To the contrary, it is fair to assume that even results from an accurate method will always differ several dozens or several hundreds from actual ridership numbers. Because they could aid in the accountability of the forecasts, the research looked at the applicability of prediction intervals to ridership forecasting. It was found that, in order to achieve a high probability for the intervals, these had to relatively wide, to the extent that anybody could guess that those forecasts. So in order for prediction intervals to have any practical use, they need to be limited in width and thus be expressed with a relatively low probability, such as 20 percent.

7.2 Limitations

A study like this, with multiple aspects and different data from various sources, is bound to have limitations. First and perhaps, foremost, this study is limited by data availability. The dependent variable, train ridership at the station level, was only available for 2004, 2005, 2006, and 2010, and for selected stations for some years in between. Ideally, these data would be available for a longer consecutive period and for all existing stations. This would allow more careful analysis of movements in ridership over time. That could give insight in maturing process of newly opened stations and would allow reliable tests of the accuracy of forecasts of the method. Additionally, it would allow the use of multiple year averages rather than ridership numbers of a single year to base the regression analyses on. This would rule out effects of bad measurements or single year outliers. That could add the reliability of the method and its forecasts. Another limitation of the dependent variable is that it is not known when the data were acquired. Seasonal and weather conditions may have had effects on the measurements, although the Dutch climate is relatively moderate.

As discussed earlier, 30 variables associated with rail ridership at the station level were identified in this research. Of these, 13 could not be included in the study because data was unavailable. These missing variables include the quality of waiting time per station, reliability of service, and road congestion. These are all thought to be closely related to train usage. But perhaps the biggest absentee has been the number of jobs within the service areas. The researcher has not had access to more disaggregated employment data. This is a major limitation, since employment is thought to be among the main drivers of train usage. As such, stations located in areas with a high job density may expect higher ridership than stations in solely residential areas. A proper, disaggregated is therefore expected to be a major contributor to better forecasts.

In this research, the train stations have been regarded from the perspective of origin stations. Partly, this is because a sufficient jobs dataset was lacking, but for another part this is due to unavailable data on the destinations of people accessing the train per individual station. The recently introduced OV chipkaart makes it possible, however, to generate data at the station level of the most important destination stations. These data help to identify the centrality of a station relative to its major destinations, and thus make it possible to explain travel behavior patterns from an individual station's perspective. It can be expected that, in the near future, the OV chipkaart will prove to be a rich source of data applicable to wide range of purposes, of which many would benefit future studies that try to account for variance in train ridership.

An important aspect of this research is the use of reference classes. It has been discussed that, in theory, there is an infinite number of reference classes for each case. From that perspective, the number of reference classes used and discussed in this study is limited. Finding the best division of the stations in order to create the ultimate reference classes will be a major puzzle. It is practically impossible to solve that puzzle in study that is limited by time and manpower. However, the study does give an indication of the way to go to achieve the best reference classes. It has been shown that smaller, more homogeneous reference classes generally lead to more accurate forecasts. On the other hand, when reference classes become too small, such as perhaps in the case of the Randstad Category 5 reference class, they become unfit for multiple regression and forecasting. Finding a balance here will require more attempts.

7.3 Suggestions for further research

As discussed in the previous section, this research has had its limitations. Future research might be able to solve these limitations, such as uncertain balance for reference classes. But at the end of this research, more questions have appeared or have remained unanswered. Firstly, it will require more research to find out whether the network buffer distance of two kilometer used here is optimal. It has been shown that for population data it is the best of the options, but perhaps the optimal distance lies in between 2000 and 3500 meters at a distance threshold not measured. Perhaps other variables require different distance for their buffers or maybe different variables require different distance buffers for different reference classes or station categories. These questions could make up a research of their own.

Another question that has not been answered is how to determine which of the forecast ridership constitutes of new train users. In other words, how does the new train station add to the total number of passengers? Especially in the case of new stations that are opened close to an existing station, a large portion of the passengers of the former will be existing train users lured away from the latter. Publically, there is not a lot of data or information available on this matter, so this would require extensive research.

Literature

Akiyama, T. & M. Okushima (2009) Analysis of railway user travel behaviour patterns of different age groups. IATSS Research 33-1, pp. 6-17.

Andriesse, H.C. & M.E.H.G. Ebben (2010) Haalbaarheid station Zandzoom. [online] [Cited on 18 October 2011] Available on the World Wide Web: <http://www.castricum.nl/document.php?m=2&fileid=28179&f=ec854befddcce319f330dd71c435c7a8&attachment=0&a=953>.

APA (2005) JAPA article calls on planners to help end inaccuracies in public project revenue forecasting. [online] [Cited on 11 October 2011] Available on the World Wide Web: <https://files.me.com/flyvbjerg/rf0t95>.

Bakker, P. & P. Zwaneveld (2009) Het belang van openbaar vervoer; De maatschappelijke effecten op een rij. The Hague: Centraal Planbureau en Kennisinstituut voor Mobiliteitsbeleid.

Beets, H. & G. Tertoolen (2009) Haalbaarheid station Barneveld Zuid. Utrecht: XTNT Experts in Traffic and Transport.

Beynen de Hoog, P. van (2004) Hoe kiest de keuzereiziger? Mogelijke determinanten van vervoerwijzekeuzegedrag onder keuzereizigers. [online] [Cited on 22 September 2011] Available on the World Wide Web: <http://www.cvs-congres.nl/cvspdfdocs/cvs0433.pdf>.

Bijl, R. van der & M. Maartens (2010) Köln Hauptbahnhof wijst Nederlands station de weg. OV magazine 2, pp. 12-15.

CBS Statline (2010) Musea; grootteklasse, bezoekersaantallen en personeel per soort museum. [online] [Cited on 3 October 2011] Available on the World Wide Web: <http://statline.cbs.nl/StatWeb/publication/?VW=T&DM=SLNL&PA=7408MUS&LA=NL>.

Cervero, R. & K. Kockelman (1997) Travel demand and the 3Ds: density, diversity and design. Transportation Research 2-3, pp. 199-219.

Cheng, E.K. (2009) A practical solution to the reference class problem. Brooklyn Law School Legal Studies Research papers Accepted Paper Series. [online] [Cited on 16 October 2011] Available on the World Wide Web: <http://ssrn.com/abstract=1357682>.

Colyvan, M., H.M. Regan & S. Ferson (2001) Is it a crime to belong to reference class? The Journal of Political Philosophy 9-2, pp. 168-181.

Currie, G., A. Ahern & A. Delbosc (2010) Exploring the drivers of light rail ridership: an empirical route level analysis of selected Australian, North American and European systems. Transportation 38, pp. 545-560.

Department of infrastructure, planning and natural resources (2003) Train users in Sydney. [online] [Cited on 22 September 2011] Available on the World Wide Web: <www.bts.nsw.gov.au/ArticleDocuments/80/train-users-2003.pdf.aspx>.

Eurlings, C. & J.C. Huizinga-Heringa (2009) Beleidsbrief nieuwe stations. Beleidsbrief 4 december. Den Haag: Ministerie van Verkeer en Waterstaat.

Flyvbjerg, B. (2006) From Nobel prize to project management: getting risks right. Project Management Journal 37-3, pp. 5-15.

Flyvbjerg, B. (2007) Cost overruns and demand shortfalls in urban rail and other infrastructure. Transportation Planning and Technology 30-1, pp. 9-30.

Flyvbjerg, B. (2008) Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. European Planning Studies 16-1, pp. 3-21.

Flyvbjerg, B. (2009) Survival of the unfittest: Why the worst infrastructure gets built-and what we can de about it. Oxford Review of Economic Policy 25-3, pp. 344-367.

Flyvbjerg, B., M.K. Skamris Holm & S.L. Buhl (2005) How (in)accurate are demand forecasts in public works projects? The case of transportation. Journal of the American Planning Association 71-2, pp. 131-146.

Givoni, M. & P. Rietveld (2007) The access journey to the railway station and its role in passengers' satisfaction with rail travel. Transport Policy 14, pp. 357-365.

Gutierrez, J. & J.C. Garcia-Palomares (2008) Distance-measure impacts on the calculation of transport areas using GIS. Environment and Planning B: Planning and design 35, pp. 480-503.

Hagen, M. van, J. Visser & M. de Gier (2005) De psychologie van de treinreiziger: een verkenning van de behoeften en drijfveren van treinreizigers. [online] [Cited on 21 September 2011] Available on the World Wide Web: <http://www.cvs-congres.nl/cvspdfdocs/cvs05_67.pdf>.

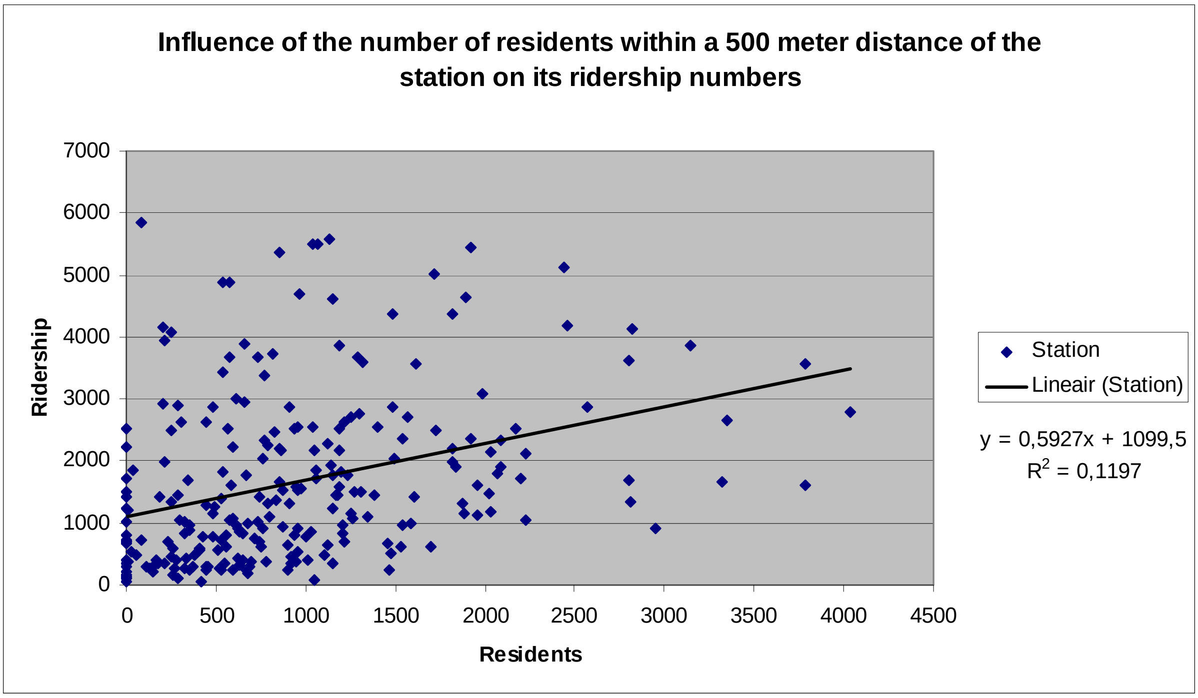
Hájek, A. (2006) The reference class problem is your problem too. [online] [Cited on 16 October 2011] Available on the World Wide Web: <http://philrsss.anu.edu.au/people-defaults/alanh/papers/rcp_your_problem_too.pdf>.

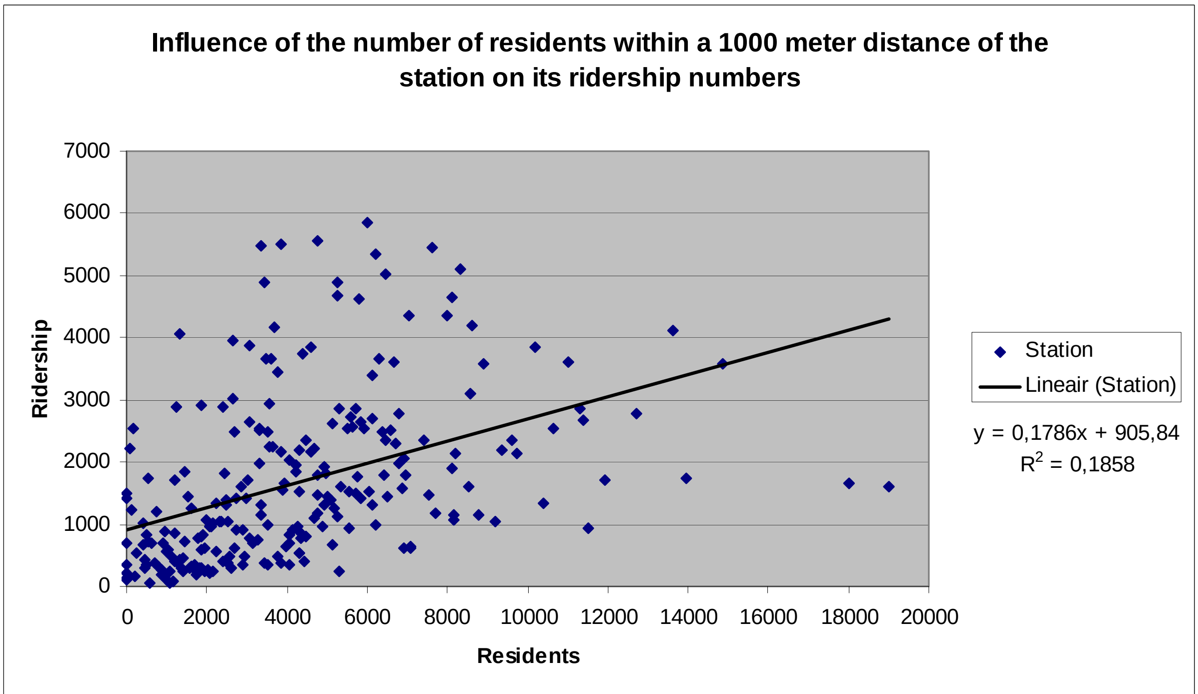
Hof, B. & C. Koopmans (2006) Toetsing prioritering stationsprojecten. Amsterdam: Seo economisch onderzoek.
Jones, L.R. & K.J. Euske (1991) Strategic misrepresentation in budgeting. Journal of Public Administration Research and Theory 4, pp. 437-460.
Keijer, M.J.N. & P. Rietveld (2000) How do people get to the railway station? The Dutch experience. Transportation Planning and Technology 23-3, pp. 215-235.
Kooij, L.P. (2009) Reactivering spoorlijn Maastricht – Aachen. Amersfoort: Arcadis Nederland BV.
Kuby, M., A. Barranda & C. Upchurch (2003) Factors influencing light-rail station boardings in the United States. Transportation Research Part A 38, pp. 223-247.
Leusden, R. van & F. Wegewijs (2009) Haalbaarheid nieuwe stations provincie Overijssel. [online] [Cited on 18 October 2011] Available on the World Wide Web: <http://www.google.nl/url?sa=t&source=web&cd=11&ved=0CBsQFjAAOAo&url=http%3A%2F%2Fwww.overijssel.nl%2Fpublish%2Fpages%2F84822%2Fedop-959335-v1-eindrapportage_nieuwe_stations.pdf&rct=j&q=kringenmethodiek&ei=DFWdToGADougOs2J9fwI&usg=AFQjCNFDr4PsOCnksjUQf3Vg-Hn1ftKD9Q>.
Levinson, D. (2011) Mischief in Minnesota: Northstar and strategic misrepresentation. [online] [Cited on 11 October 2011] Available on the World Wide Web: <http://blog.lib.umn.edu/levin031/transportationist/2011/04/mischief-in-minnesota-northsta.html>.
Lovallo, D. & D. Kahneman (2003) Delusions of success: How optimism undermines executives' decisions. Harvard Business Review, July, pp. 56-63.
Nanninga, H. & M. Woolthuis (2004) Behoud en versterking stations in Velsen; Onderzoek naar de potentie van de stations Driehuis, Santpoort Noord en Santpoort Zuid en de effecten van een gewijzigde lijnvoering van bus en trein op en rond de contractsectorspoorlijn Haarlem – Uitgeest. Inno-V Adviseurs.
NS (2011) Over onze klanten. [online] [Cited on 18 August 2011] Available on the World Wide Web: <http://www.ns.nl/over-ns/wat-doen-wij/ontdek-ns/wetenswaardigheden/over-onze-klanten.html>.
NY Times (2011) A congressman's pet project; A railroad's boon. June 27.
Mackett, R.L. & M. Edwards (1997) The impact of new urban public transport systems: will the expectations be met? Transportation Research A 32-4, pp. 231-245.
Miller, H.J. (2004) Tobler's first law and spatial analysis. Annals of the Association of American Geographers 94-2, pp. 284-289.
Ministerie van Verkeer en Waterstaat (2009) MIRT Projectenboek; Kleine stations. [online] [Cited on 16 August 2011] Available on the World Wide Web: <http://mirt2009.mirtprojectenboek.nl/images/515_tcm285-228660.pdf>.
Oort, N. van (2011) Service reliability and urban public transport design. Delft: Trail Research School.
Projectteam Toepassing Norm ProRail (2009) Toepassing norm nieuwe in- en uitstappers bij nieuwe stations; Inventarisatie. Utrecht: ProRail.
Spit, T. & P. Zoete (2005) De achterkant van de planologie; Een inleiding in ruimtelijke ordening en planologie. Den Haag: SDU Uitgevers.
Stuurgroep Toekomstonderzoek en strategisch Omgevingsbeleid (2000) Terugblik op toekomstverkenningen. Den Haag: WRR.
Sung, H. & J. Oh (2011) Transit-oriented development in a high-density city: Identifying its association with transit ridership in Seoul, Korea. Cities 28, pp. 70-82.
Tal, G. (2008) Overestimation reduction in forecasting telecommuting as a TDM policy. Davis: University of California.
Treinreiziger.nl (2010) Punctualiteitscijfers NS door de jaren heen. [online] [Cited on 3 October 2011] Available on the World Wide Web: <http://www.treinreiziger.nl/kennisnet/punctualiteit/punctualiteitscijfers/punctualiteiscijfers_ns_door_de_jaren_heen>.
Tobler, W. (2004) On the first law of geography. Annals of the Association of American Geographers 94-2, pp. 304-310.
Tweede Kamer der Staten-Generaal (2008) Nieuwe stations, nieuwe passagiers. The Hague: SDU Uitgevers.
Vanston, J.H. & L.K. Vanston (2004) Testing the tea leaves: Evaluating the validity of forecasts. Research-Technology Management, 47(5), pp. 33-39.
Verburg, T., M. Dijst & T. Schwanen (2005) Leef- en mobiliteitstijlen Stedenbaan. Utrecht: Faculty of Geosciences.
Vegouwen, J. & J. Baggen (2004) Een integraal afwegingskader voor nieuwe treinstations; Breda Oost als voorbeeld. [online] [Cited on 16 August 2011] Available on the World Wide Web: <http://www.cvs-congres.nl/cvspdfdocs/cvs0472.pdf>.

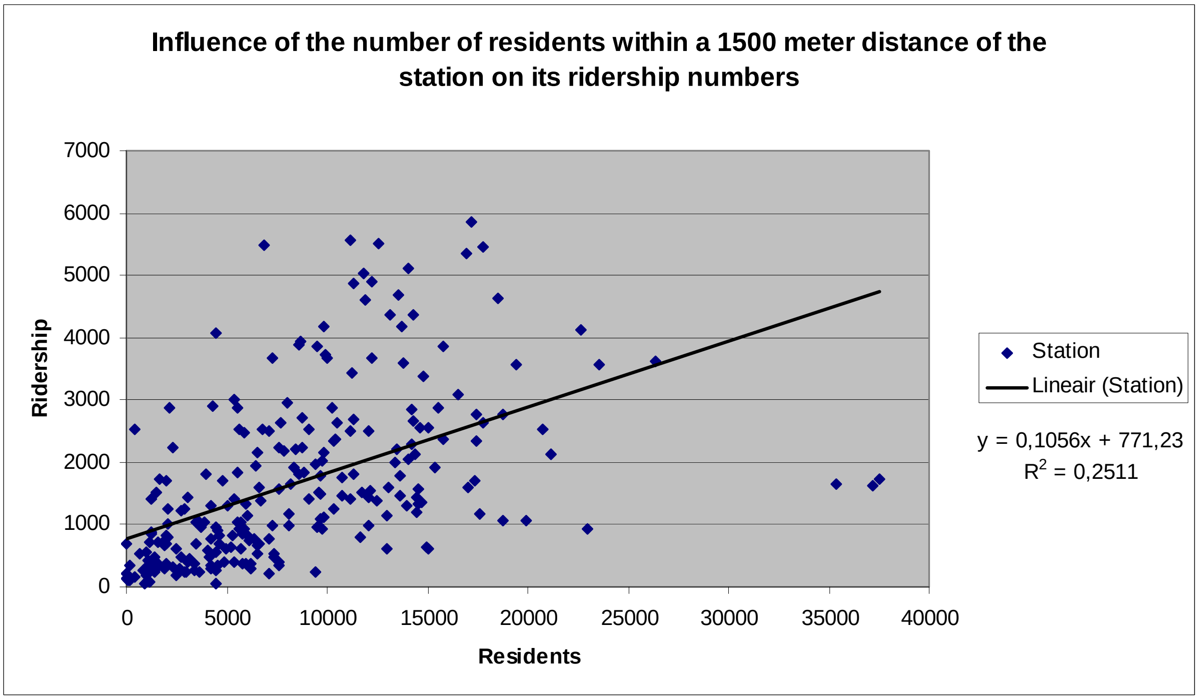
Vocht, A. de (2005a) Syllabus statistiek voor de opleiding Sosiale Geografie en Planologie. Utrecht: Faculteit Geowetenschappen.
Vocht, A. de (2005b) Basishandboek SPSS voor windows. Utrecht: Bijleveld Press.
Vos, A. de (2009) Rijkswaterstaat en NS vinden elkaar op de A2. OV magazine 3, 12-14.
Vuuren, D. van (2002) The market for passenger transport by train. Amsterdam: Vrije Universiteit.
Witteveen + Bos (2010) Quickscan haltes Baexum en Haelen. [online] [Cited on 18 October 2011] Available on the World Wide Web: <http://www.witteveenbos.nl/nl/projecten-verkeer-en-vervoer/project/353-quickscan-haltes-baexem-en-haelen>.

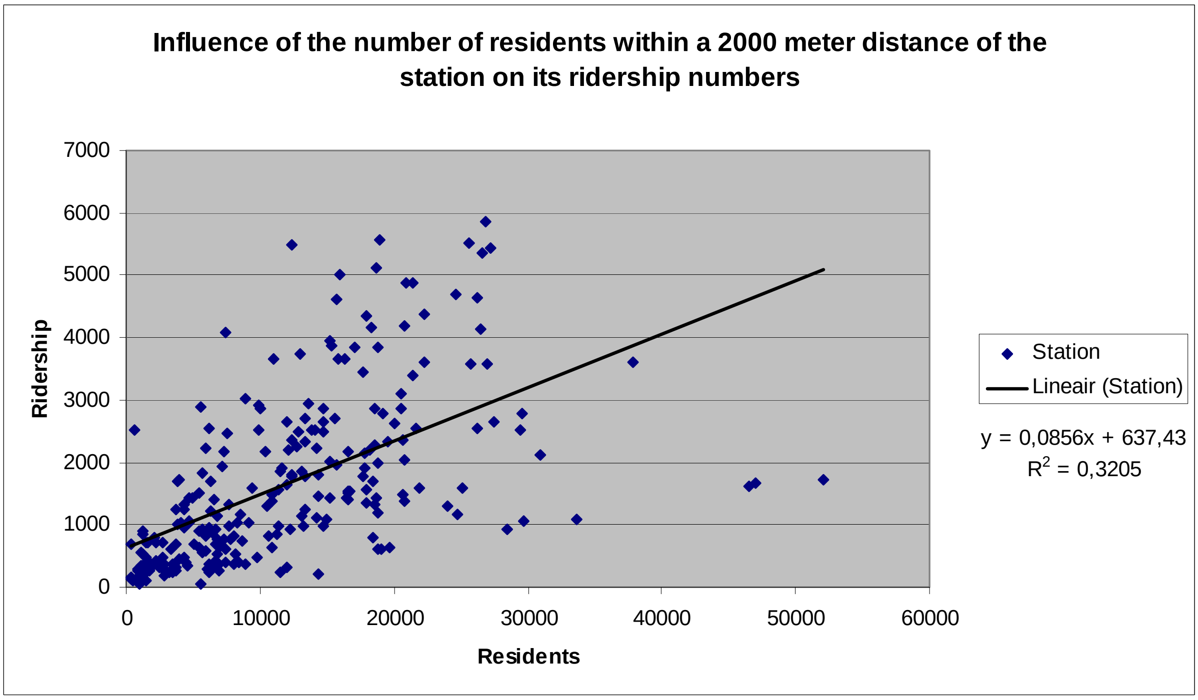
Appendix 1: Extent of service areas

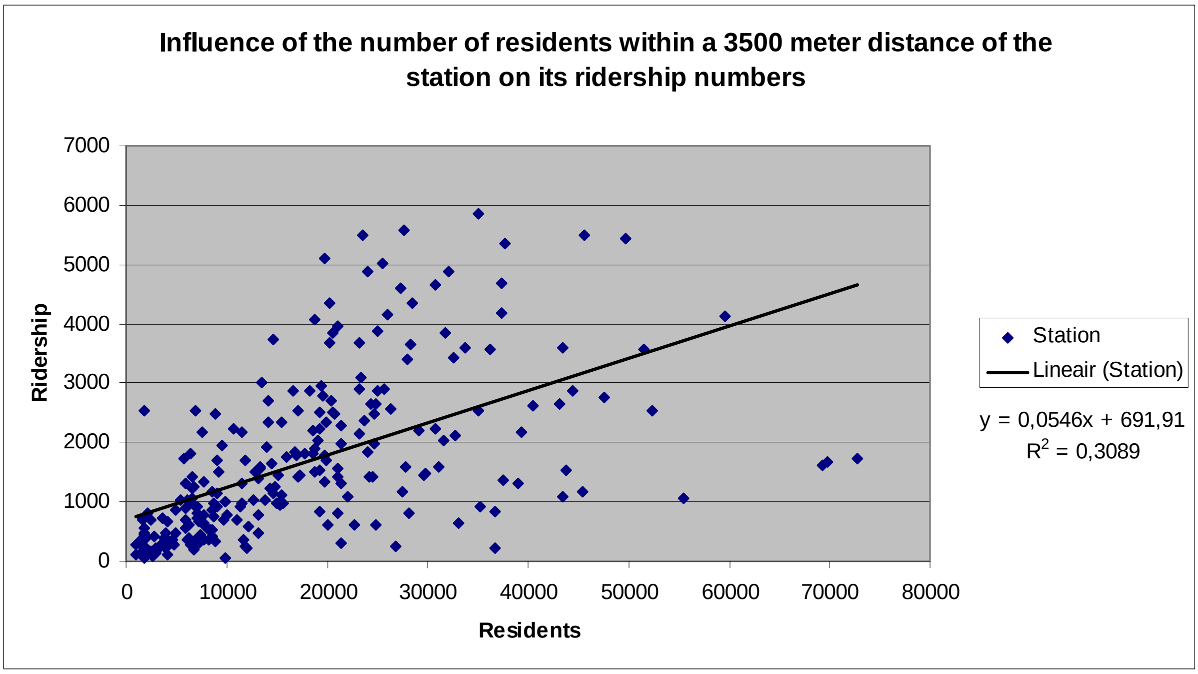
Range	R^2
0-500	0.1197
0-1000	0.1858
0-1500	0.2511
0-2000	0.3205
0-3500	0.3089
0-5000	0.2650



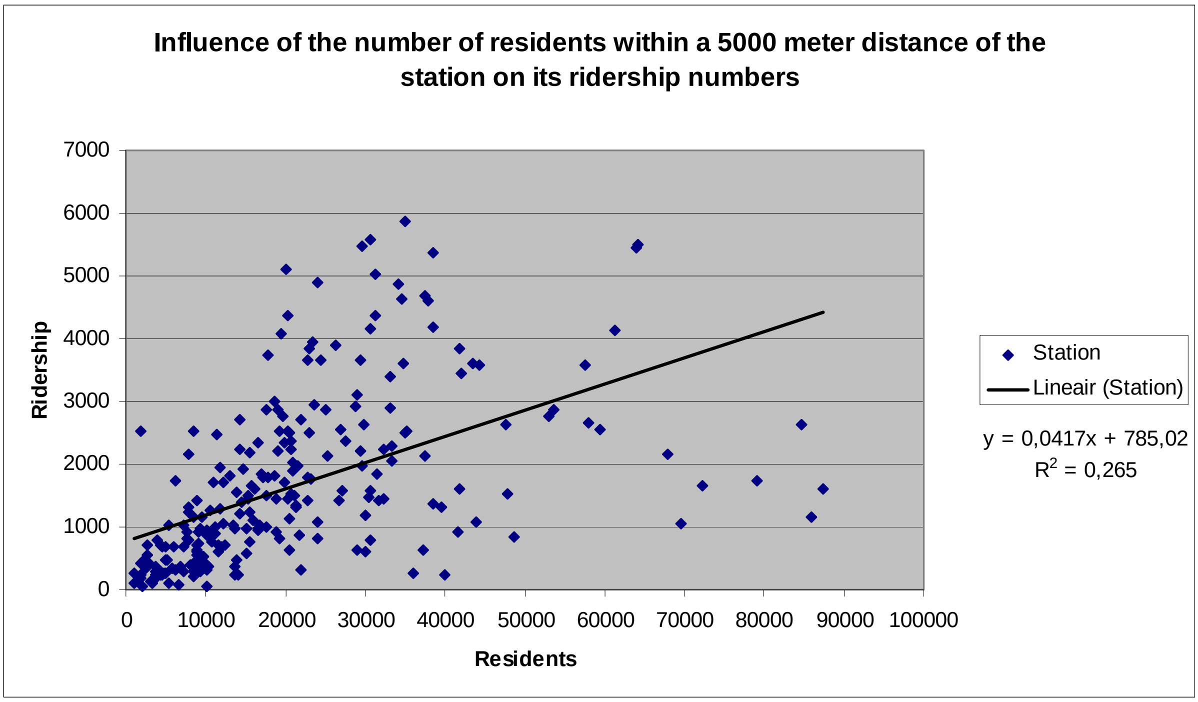



Influence of the number of residents within a 1500 meter distance of the station on its ridership numbers
Ridership
7000
6000
5000
4000
3000
2000
1000
0
0
5000
10000
15000
20000
25000
30000
35000
40000
Residents
Station
Lineair (Station)
y = 0,1056x + 771,23
R^2 = 0,2511


Influence of the number of residents within a 2000 meter distance of the station on its ridership numbers
Ridership
7000
6000
5000
4000
3000
2000
1000
0
0
10000
20000
30000
40000
50000
60000
Residents
Station
Lineair (Station)
y = 0,0856x + 637,43
R^2 = 0,3205



Influence of the number of residents within a 3500 meter distance of the station on its ridership numbers
Ridership
7000
6000
5000
4000
3000
2000
1000
0
0
10000
20000
30000
40000
50000
60000
70000
80000
Residents
Station
Lineair (Station)
y = 0,0546x + 691,91
R^2 = 0,3089



Influence of the number of residents within a 5000 meter distance of the station on its ridership numbers
Ridership
7000
6000
5000
4000
3000
2000
1000
0
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
Residents
Station
Lineair (Station)
y = 0,0417x + 785,02
R^2 = 0,265

Appendix 2: SPSS output for the All Stations reference class

Table A2.1: Part of the correlation matrix from SPSS showing multicollinearity between total of stations accessible variables.

		N2006	TOT_60	IC_60	STOP_60	TOT_30
N2006	Pearson Correlation	1	,514**	,524**	,485**	,563**
	Sig. (2-tailed)		,000	,000	,000	,000
	N	296	296	296	296	296
TOT_60	Pearson Correlation	,514**	1	,922**	,986**	,756**
	Sig. (2-tailed)	,000		,000	,000	,000
	N	296	296	296	296	296
IC_60	Pearson Correlation	,524**	,922**	1	,844**	,658**
	Sig. (2-tailed)	,000	,000		,000	,000
	N	296	296	296	296	296
STOP_60	Pearson Correlation	,485**	,986**	,844**	1	,762**
	Sig. (2-tailed)	,000	,000	,000		,000
	N	296	296	296	296	296
TOT_30	Pearson Correlation	,563**	,756**	,658**	,762**	1
	Sig. (2-tailed)	,000	,000	,000	,000	
	N	296	296	296	296	296
IC_30	Pearson Correlation	,492**	,666**	,711**	,614**	,708**
	Sig. (2-tailed)	,000	,000	,000	,000	,000
	N	296	296	296	296	296
STOP_30	Pearson Correlation	,480**	,641**	,493**	,675**	,935**
	Sig. (2-tailed)	,000	,000	,000	,000	,000
	N	296	296	296	296	296
CROSSINGS	Pearson Correlation	,424**	,341**	,383**	,307**	,278**
	Sig. (2-tailed)	,000	,000	,000	,000	,000
	N	296	296	296	296	296
BUILTUP_P	Pearson Correlation	,409**	,309**	,341**	,280**	,322**
	Sig. (2-tailed)	,000	,000	,000	,000	,000
	N	296	296	296	296	296
CII_RES_RATE	Pearson Correlation	,049	,033	,039	,029	,021
	Sig. (2-tailed)	,403	,573	,502	,623	,714

**. Correlation is significant at the 0.01 level (2-tailed).

TOT_60 and STOP_30 are excluded due to multicollinearity (see table A2.1). Other excluded variables include POP_15T65, HH_TOTAL, ONEP_HH, MP_HH_NC, MP_HH_WC, and CAR_OWN (see table A2.2).

Table A2.2: Part of the correlation matrix from SPSS showing multicollinearity between several population variables.

		CII_RES_RATE	POP_TOTAL	POP_15T65
CII_RES_RATE	N	296	296	296
POP_TOTAL	Pearson Correlation	-,131*	1	,996**
	Sig. (2-tailed)	,024		,000
	N	296	296	296
POP_15T65	Pearson Correlation	-,121*	,996**	1
	Sig. (2-tailed)	,038	,000	
	N	296	296	296
POP_ALLOCH	Pearson Correlation	-,002	,776**	,790**
	Sig. (2-tailed)	,969	,000	,000
	N	296	296	296
HH_TOTAL	Pearson Correlation	-,106	,980**	,986**
	Sig. (2-tailed)	,070	,000	,000
	N	296	296	296
ONEP_HH	Pearson Correlation	-,059	,885**	,906**
	Sig. (2-tailed)	,315	,000	,000
	N	296	296	296
MP_HH_NC	Pearson Correlation	-,145*	,977**	,969**
	Sig. (2-tailed)	,013	,000	,000
	N	296	296	296
MP_HH_WC	Pearson Correlation	-,142*	,971**	,960**
	Sig. (2-tailed)	,014	,000	,000
	N	296	296	296
CAR_OWN	Pearson Correlation	-,149*	,927**	,911**
	Sig. (2-tailed)	,010	,000	,000
	N	296	296	296

Correlations not displayed here are all below the threshold value of 0.9. The corresponding variables can thus be included in the MR analysis.

Table A2.3: Casewise diagnostics showing three outliers with a standardized residual of >4.0

Case Number	Std. Residual	N2006	Predicted Value	Residual
17	8,467	21153	4548,72	16604,282
271	3,809	12465	4994,35	7470,649
273	-4,918	4126	13770,09	-9644,086
280	6,052	39555	27686,97	11868,025

Table A2.4: Casewise diagnostics showing two outliers with a standardized residual of >4.0

Case Number	Std. Residual	N2006	Predicted Value	Residual
226	3,640	12131	7556,75	4574,252
254	4,959	12574	6342,88	6231,121
270	6,571	12465	4208,19	8256,810

The first two MR analyses both yield outliers with standardized residual of more than 0.4 (see table A2.3 and table A2.4). The outliers include, but are not limited to, Amsterdam Sloterdijk, Rotterdam Blaak, and Schiedam Centrum. These are removed and a new attempt is subsequently made. The third run does not yield outliers.

Table A2.5: Coefficients of the variables included in the regression function

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
11	(Constant)	-1260,123	232,469		-5,421	,000
	FREQUENCY_oS	247,409	38,326	,322	6,455	,000
	CROSSINGS	2,113	,498	,222	4,239	,000
	BTM_TOTAL	42,169	4,723	,343	8,928	,000
	CATEGORY_5	-1259,384	163,340	-,265	-7,710	,000
	IC_60	101,574	17,782	,236	5,712	,000
	TYPE_SCORE	650,750	201,085	,102	3,236	,001
	BUILTUP_P	20,721	5,500	,192	3,767	,000
	TRANSFER	1218,108	289,088	,140	4,214	,000
	ONEP_HH	-,085	,037	-,113	-2,327	,021
	URBANIZATION_4	-347,339	151,451	-,071	-2,293	,023
	TOT_30	-68,584	33,987	-,112	-2,018	,045

The regression function can be derived from the unstandardized B coefficients (see table A2.5). The function is depicted below, where Y is the stations' ridership number.

Y = -1260.1 + 247.4 * FREQUENCY_oS + 2.1 * CROSSINGS + 42.2 * BTM_TOTAL -1259.4 * CATEGORY_5 + 101.6 * IC_60 + 650.8 * TYPE_SCORE + 20.7 * BUILTUP_P + 1218.1 * TRANSFER -0.085 * ONEP_HH -347.3 * URBANIZATION_4 -68.6 * TOT_30

The model is significant at the 0.000 percent level (see table A2.6).

Table A2.6: ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
11	Regression	1,030E9	11	93610540,842	83,934	,000[k]
	Residual	3,078E8	276	1115284,801		
	Total	1,338E9	287			

The model's coefficient of determination is 0.770, which means the model explains 77 percent of the variance in stations' ridership numbers (see table A2.7).

Table A2.7: Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
11	,877[k]	,770	,761	1056,070

The model does not include the variable POP_TOTAL due to insignificance. Because there is a theoretical basis to include it anyway, it has been tried to do so with the Enter method instead of the Stepwise method (see table A2.7).

Table A2.8: Coefficients for a model including the insignificant variable POP_TOTAL

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1229,811	234,184		-5,251	,000
	IC_60	100,428	17,997	,233	5,580	,000
	TOT_30	-68,441	33,924	-,112	-2,017	,045
	CROSSINGS	1,976	,622	,208	3,176	,002
	BUILTUP_P	20,004	5,597	,187	3,574	,000
	POP_TOTAL	,008	,022	,039	,365	,715
	ONEP_HH	-,102	,060	-,136	-1,710	,088
	BTM_TOTAL	42,358	4,707	,344	8,999	,000
	FREQUENCY_oS	246,014	38,568	,320	6,379	,000
	URBANIZATION_4	-343,450	150,183	-,071	-2,287	,023
	TRANSFER	1236,588	293,050	,142	4,220	,000
	TYPE_SCORE	650,099	200,458	,102	3,243	,001
	CATEGORY_5	-1258,294	162,792	-,264	-7,729	,000

Not only does table A2.8 show that POP_TOTAL is far from significant, but including it in the model also causes ONEP_HH to become insignificant at the 95 percent.

Figure A2.1: Histogram of the normal distribution of the residuals

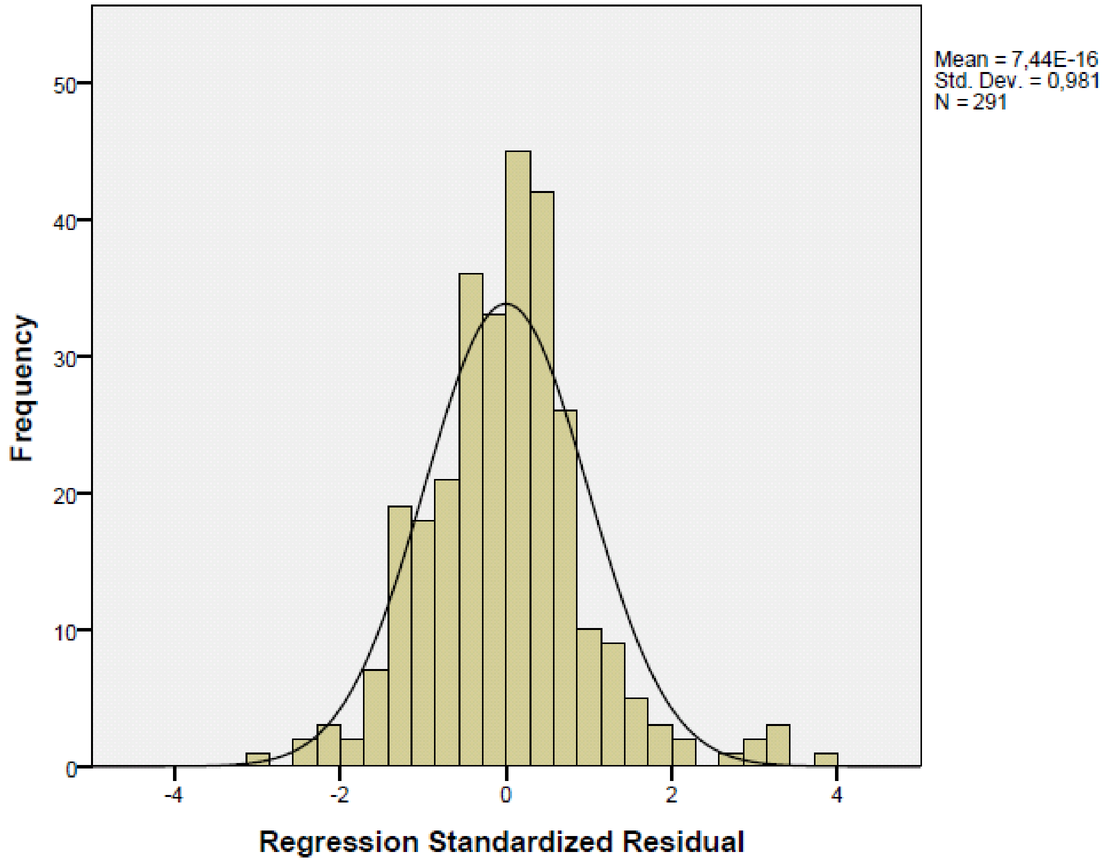


Figure A2.2: Normal probability plot of the standardized residual

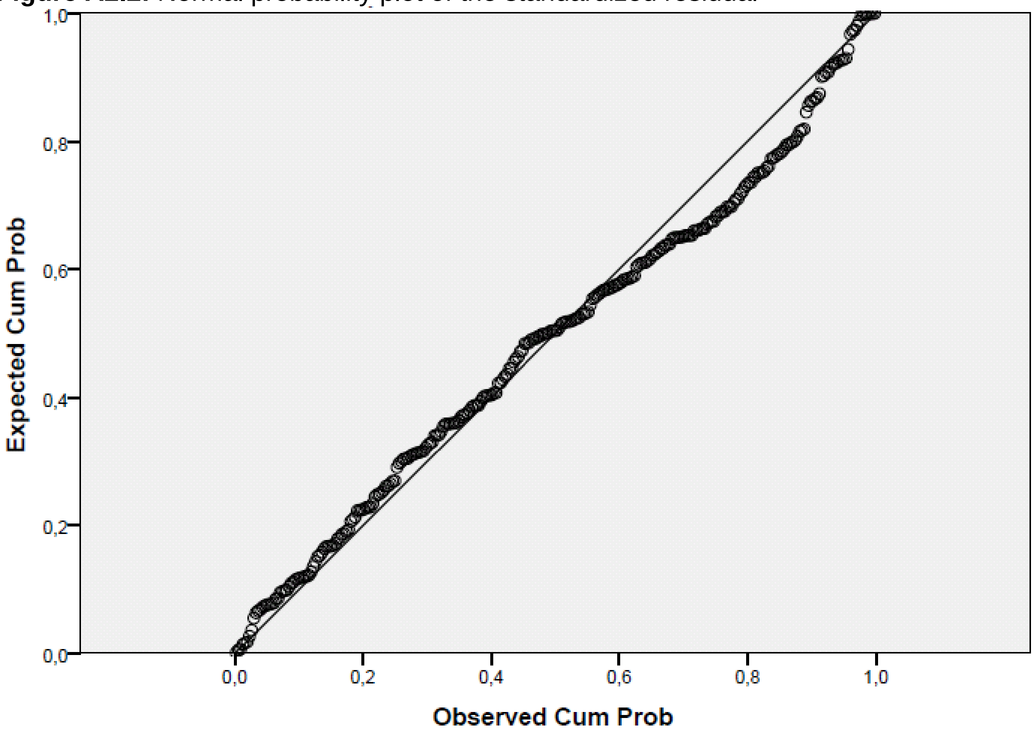
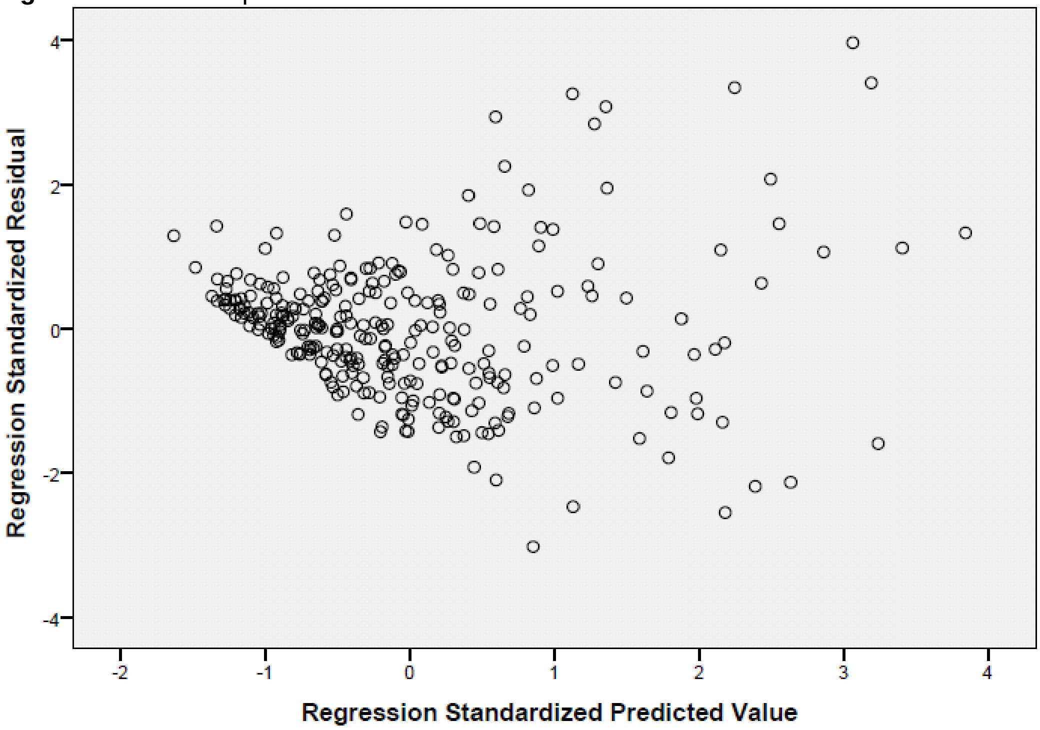


Figure A2.3: Scatterplot of the distribution of the standardized residuals



Appendix 3: SPSS output for the Randstad Stations reference class

Table A3.1: Part of the correlation matrix from SPSS

		N2006	TOT_60	IC_60	STOP_60	TOT_30
N2006	Pearson Correlation	1	,497**	,490**	,479**	,641**
	Sig. (2-tailed)		,000	,000	,000	,000
	N	79	79	79	79	79
TOT_60	Pearson Correlation	,497**	1	,926**	,989**	,672**
	Sig. (2-tailed)	,000		,000	,000	,000
	N	79	79	79	79	79
IC_60	Pearson Correlation	,490**	,926**	1	,858**	,588**
	Sig. (2-tailed)	,000	,000		,000	,000
	N	79	79	79	79	79
STOP_60	Pearson Correlation	,479**	,989**	,858**	1	,678**
	Sig. (2-tailed)	,000	,000	,000		,000
	N	79	79	79	79	79
TOT_30	Pearson Correlation	,641**	,672**	,588**	,678**	1
	Sig. (2-tailed)	,000	,000	,000	,000	
	N	79	79	79	79	79
IC_30	Pearson Correlation	,572**	,687**	,683**	,660**	,830**
	Sig. (2-tailed)	,000	,000	,000	,000	,000
	N	79	79	79	79	79
STOP_30	Pearson Correlation	,586**	,565**	,448**	,588**	,949**
	Sig. (2-tailed)	,000	,000	,000	,000	,000
	N	79	79	79	79	79
CROSSINGS	Pearson Correlation	,325**	,362**	,305**	,369**	,370**
	Sig. (2-tailed)	,003	,001	,006	,001	,001
	N	79	79	79	79	79
BUILTUP_P	Pearson Correlation	,365**	,249*	,202	,257*	,492**
	Sig. (2-tailed)	,001	,027	,074	,022	,000
	N	79	79	79	79	79
CII_RES_RATE	Pearson Correlation	,347**	,227*	,240*	,212	,328**
	Sig. (2-tailed)	,002	,044	,033	,060	,003

STOP_60 and STOP_30 are excluded due to multicollinearity with the variables TOT_60 and TOT_30 respectively. The latter two are included at the cost of the former two because the latter show higher correlation with ridership. Other variables excluded, due to multicollinearity with POP_TOTAL, are POP_15T65, HH_TOTAL, ONEP_HH, MP_HH_NC, MP_HH_WC, and CAR_OWN (not shown).

Table A3.2: Casewise diagnostics showing no outliers with a standardized residual of >4.0

Case Number	Std. Residual	N2006	Predicted Value	Residual
63	3,259	39555	30772,89	8782,111

Table A3.3: Coefficients of the variables included in the regression function

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
4	(Constant)	-3624,606	886,655		-4,088	,000
	FREQUENCY_oS	745,704	60,414	,799	12,343	,000
	BTM_BINAIR	2490,585	827,549	,182	3,010	,004
	CATEGORY_5	-1937,632	670,693	-,190	-2,889	,005
	JOBS_MUN	,005	,002	,145	2,107	,039

As shown by table A3.3, the first run in SPSS did not bring forward any outliers. The MR analysis gave the resulting coefficients as presented in table A3.3, which lead to the following function:

Y = -3624.6 + 745.7 * FREQUENCY_oS + 2490.6 * BTM_BINAIR -1937.6 * CATEGORY_5 + 0.005 * JOBS_MUN

The model is significant at the 0.000 percent level (see table A3.4) with a coefficient of determination of 0.736, explaining 73.6 percent of the variance in the dependent variable. Judging by the beta coefficients from table A3.3, FREQUENCY_oS is the most important variable is this function.

Table A3.4: ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
4	Regression	1,498E9	4	3,745E8	51,558	,000[d]
	Residual	5,375E8	74	7263492,916		
	Total	2,035E9	78			

Table A3.5: Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
4	,858[d]	,736	,722	2695,087

This model does not include the variable POP_TOTAL, while there is strong theoretical evidence that it should. Including it, however, hardly adds anything to the model's accuracy. Furthermore, the variable is highly insignificant, as shown by table A3.6.

Table A3.6: Model coefficients including the variable POP_TOTAL

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-3544,290	912,001		-3,886	,000
	FREQUENCY_oS	750,435	61,793	,804	12,144	,000
	BTM_BINAIR	2562,850	849,876	,187	3,016	,004
	CATEGORY_5	-1878,162	689,228	-,185	-2,725	,008
	JOBS_MUN	,005	,002	,155	2,114	,038
	POP_TOTAL	-,012	,028	-,031	-,419	,676

Figure A3.1: Histogram of the normal distribution of the residuals

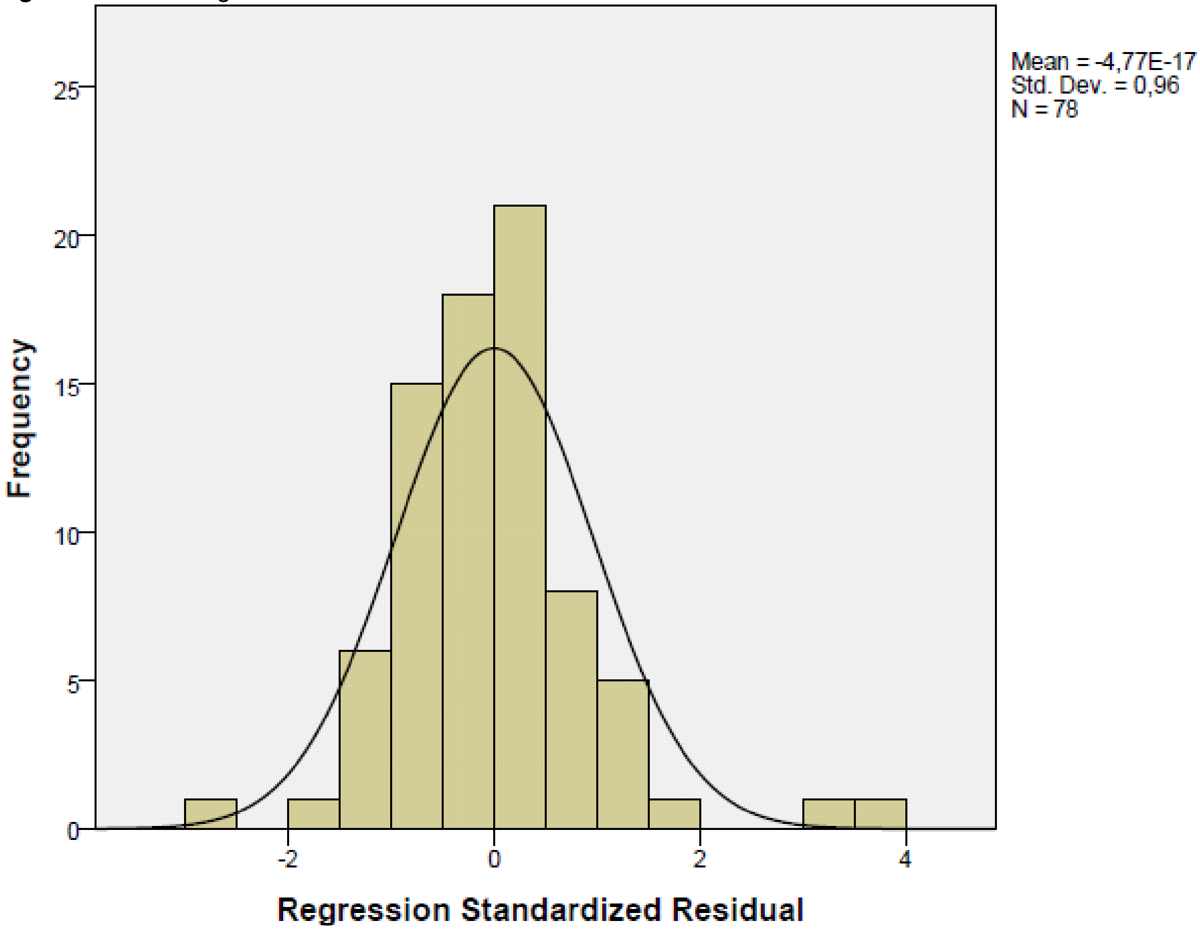


Figure A3.2: Normal probability plot of the standardized residual

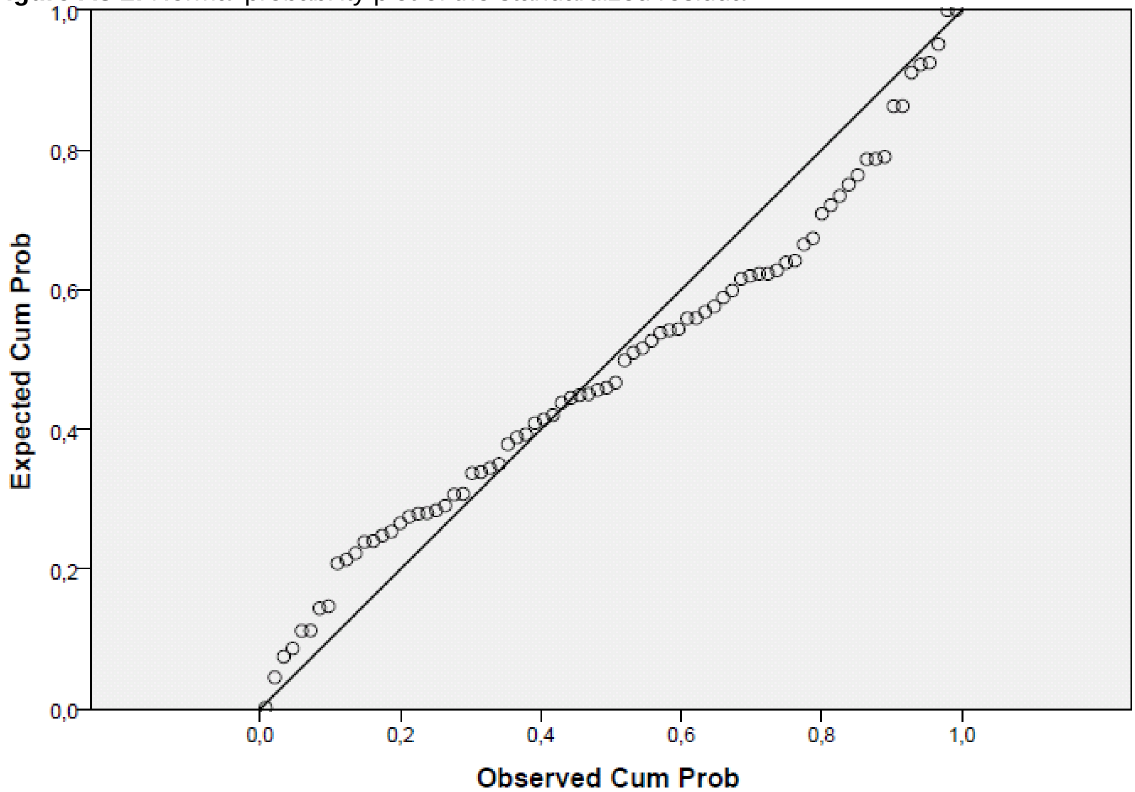
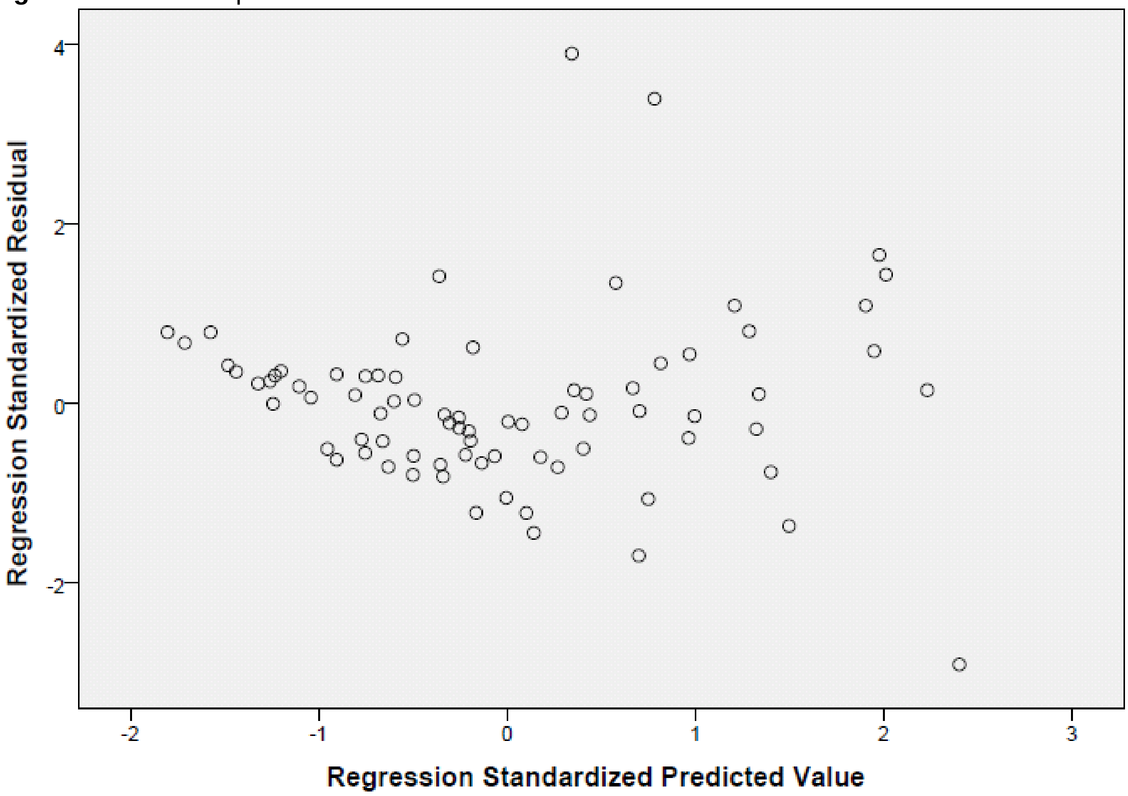


Figure A3.3: Scatterplot of the distribution of the standardized residuals



Appendix 4: SPSS output for the Non Randstad stations reference class

Table A4.1: Part of the correlation matrix from SPSS

		N2006	TOT_60	IC_60	STOP_60	TOT_30
N2006	Pearson Correlation	1	,321**	,387**	,254**	,224**
	Sig. (2-tailed)		,000	,000	,000	,001
	N	217	217	217	217	217
TOT_60	Pearson Correlation	,321**	1	,797**	,967**	,670**
	Sig. (2-tailed)	,000		,000	,000	,000
	N	217	217	217	217	217
IC_60	Pearson Correlation	,387**	,797**	1	,616**	,455**
	Sig. (2-tailed)	,000	,000		,000	,000
	N	217	217	217	217	217
STOP_60	Pearson Correlation	,254**	,967**	,616**	1	,681**
	Sig. (2-tailed)	,000	,000	,000		,000
	N	217	217	217	217	217
TOT_30	Pearson Correlation	,224**	,670**	,455**	,681**	1
	Sig. (2-tailed)	,001	,000	,000	,000	
	N	217	217	217	217	217
IC_30	Pearson Correlation	,145*	,457**	,618**	,333**	,376**
	Sig. (2-tailed)	,033	,000	,000	,000	,000
	N	217	217	217	217	217
STOP_30	Pearson Correlation	,178**	,521**	,218**	,588**	,913**
	Sig. (2-tailed)	,009	,000	,001	,000	,000
	N	217	217	217	217	217

Table A4.2: Part of the correlation matrix from SPSS showing hgh correlation between POP_TOTAL and CROSSINGS and BUILTUP_P

		CII_RES_RATE	POP_TOTAL	POP_15T65
CROSSINGS	Pearson Correlation	-,146*	,906**	,883**
	Sig. (2-tailed)	,032	,000	,000
	N	217	217	217
BUILTUP_P	Pearson Correlation	-,014	,847**	,828**
	Sig. (2-tailed)	,836	,000	,000
	N	217	217	217

The variables STOP_60 and STOP_30 are excluded because the multicollinearating variables TOT_60 and TOT_30 show better correlations with the dependent variable (see table A4.1). The variable CROSSINGS is excluded because the Pearson Correlation value exceeds 0.9 (see table A4.2). Other variables excluded for multicollinearity include POP_15T65, HH_TOTAL, MP_HH_NC, MP_HH_WC, and CAR_OWN.

Table A4.3: Casewise diagnostics of the first attempt

Case Number	Std. Residual	N2006	Predicted Value	Residual
16	2,141	7724	4563,63	3160,370
17	11,209	21153	4608,22	16544,777
29	2,074	7569	4506,92	3062,083
183	2,383	6380	2862,46	3517,537
199	-2,009	599	3564,40	-2965,397

The first run of the model in SPSS shows an extreme outlier in the case of Duiven station, with a standardized residual of >11 (see table A4.3). After the removal of this case, a new attempt is made. Again, the model provides outliers with a value exceeding 4 (see table A4.4). These two cases are subsequently removed.

Table A4.4: Casewise diagnostics of the second attempt

Case Number	Std. Residual	N2006	Predicted Value	Residual
16	4,081	7724	4234,04	3489,955
18	-2,181	622	2487,33	-1865,332
28	4,028	7569	4123,98	3445,017
30	2,704	5771	3458,56	2312,435
101	3,024	5486	2899,58	2586,418
173	-2,186	1055	2924,31	-1869,315
182	3,241	6380	3608,45	2771,550
198	-3,401	599	3507,62	-2908,625

Table A4.5: Casewise diagnostics of the third attempt

Case Number	Std. Residual	N2006	Predicted Value	Residual
10	-2,178	1592	3300,56	-1708,562
25	2,062	4876	3258,17	1617,828
28	3,444	5771	3068,73	2702,267
99	3,833	5486	2478,90	3007,097
171	-2,637	1055	3123,50	-2068,497
175	-2,245	1159	2920,00	-1761,004
180	3,413	6380	3702,14	2677,859
196	-3,355	599	3231,51	-2632,513

The third MR run for this reference class does not show any outliers (see table A4.5). The model is thus accepted.

The model contains eight variables, of which POP_TOTAL is relatively the most influential. All are significant at the 95 percent level (see table A4.6). The coefficients lead to the following function:

Y = -592.8 + 0.101 * POP_TOTAL + 214.0 * FREQUENCY_oS -1040.1 * CATEGORY_5 + 44.4 * BTM_TOTAL -4194.2 * URBANIZATION_1 + 373.2 * TYPE_SCORE + 53.8 * IC_60 -62.0 * TOT_30

Table A4.6: Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
8	(Constant)	-592,757	169,764		-3,492	,001
	POP_TOTAL	,101	,009	,568	11,548	,000
	FREQUENCY_oS	214,035	38,099	,311	5,618	,000
	CATEGORY_5	-1040,114	144,716	-,294	-7,187	,000
	BTM_TOTAL	44,396	6,558	,283	6,769	,000
	URBANIZATION_1	-4194,249	901,865	-,199	-4,651	,000
	TYPE_SCORE	373,225	178,784	,086	2,088	,038
	IC_60	53,847	21,427	,120	2,513	,013
	TOT_30	-61,992	30,199	-,111	-2,053	,041

Table A4.7: ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
8	Regression	3,198E8	8	39970848,218	64,938	,000[h]
	Residual	1,243E8	202	615525,295		
	Total	4,441E8	210			

The model is significant at the 0.000 percent level (see table A4.7). The R^2 is 0.72 (see table A4.8).

Table A4.8: Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
8	,849[h]	,720	,709	784,554

Figure A4.1: Histogram of the normal distribution of the residuals

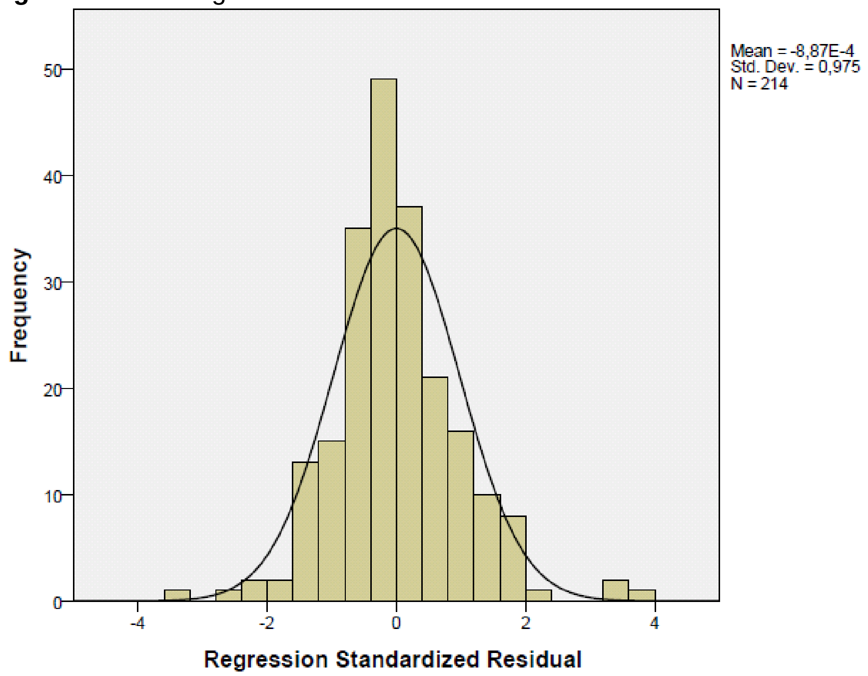


Figure A4.2: Normal probability plot of the standardized residual

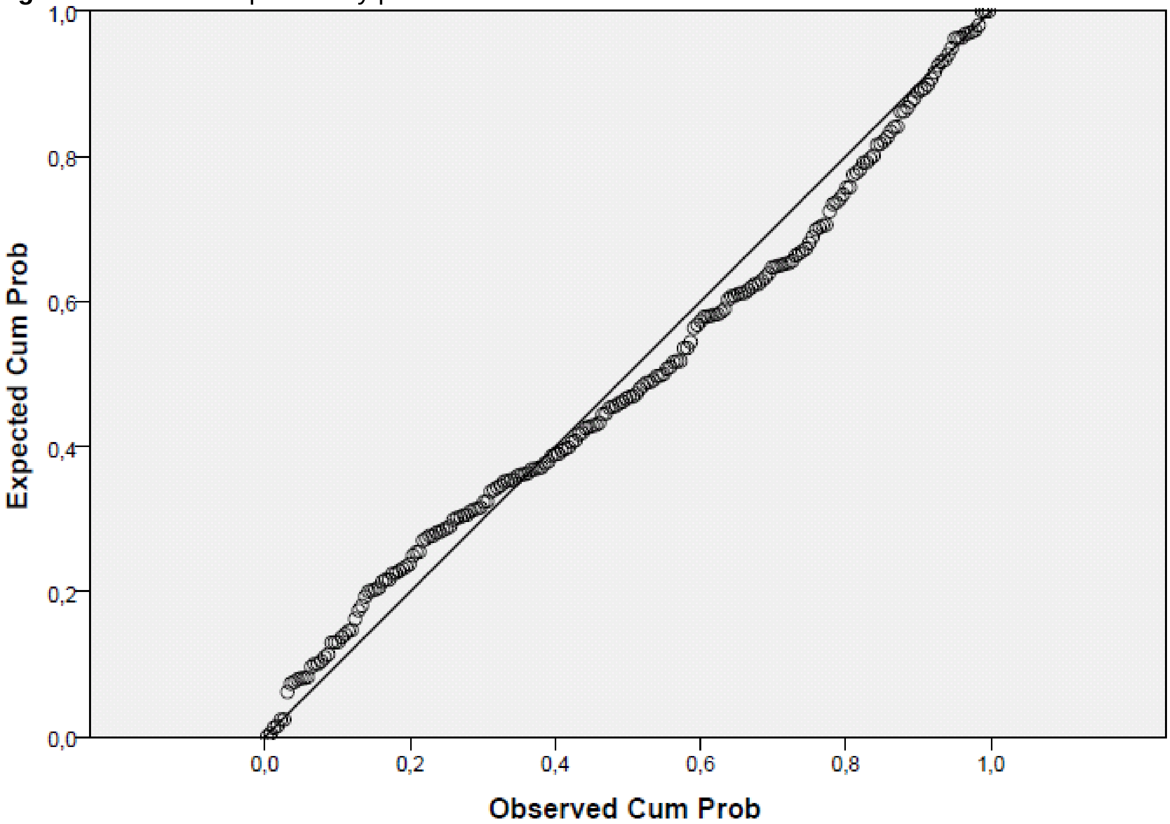


Figure A4.3: Scatterplot of the distribution of the standardized residuals

Appendix 5: SPSS output for the Category 4 stations reference class

Table A5.1: Part of the correlation matrix from SPSS

		N2006	TOT_60	IC_60	STOP_60	TOT_30
N2006	Pearson Correlation	1	,510**	,503**	,483**	,420**
	Sig. (2-tailed)		,000	,000	,000	,000
	N	124	124	124	124	124
TOT_60	Pearson Correlation	,510**	1	,906**	,983**	,668**
	Sig. (2-tailed)	,000		,000	,000	,000
	N	124	124	124	124	124
IC_60	Pearson Correlation	,503**	,906**	1	,813**	,543**
	Sig. (2-tailed)	,000	,000		,000	,000
	N	124	124	124	124	124
STOP_60	Pearson Correlation	,483**	,983**	,813**	1	,684**
	Sig. (2-tailed)	,000	,000	,000		,000
	N	124	124	124	124	124
TOT_30	Pearson Correlation	,420**	,668**	,543**	,684**	1
	Sig. (2-tailed)	,000	,000	,000	,000	
	N	124	124	124	124	124
IC_30	Pearson Correlation	,283**	,496**	,520**	,458**	,552**
	Sig. (2-tailed)	,001	,000	,000	,000	,000
	N	124	124	124	124	124
STOP_30	Pearson Correlation	,366**	,561**	,404**	,597**	,926**
	Sig. (2-tailed)	,000	,000	,000	,000	,000
	N	124	124	124	124	124

Table A5.2: Part of the correlation matrix from SPSS

		CII_RES_RATE	POP_TOTAL	POP_15T65
HH_TOTAL	Pearson Correlation	-,215*	,982**	,975**
	Sig. (2-tailed)	,017	,000	,000
	N	124	124	124
ONEP_HH	Pearson Correlation	-,127	,863**	,846**
	Sig. (2-tailed)	,159	,000	,000
	N	124	124	124
MP_HH_NC	Pearson Correlation	-,220*	,974**	,966**
	Sig. (2-tailed)	,014	,000	,000
	N	124	124	124
MP_HH_WC	Pearson Correlation	-,272**	,983**	,988**
	Sig. (2-tailed)	,002	,000	,000
	N	124	124	124
CAR_OWN	Pearson Correlation	-,259**	,985**	,977**
	Sig. (2-tailed)	,004	,000	,000
	N	124	124	124

Variables excluded due to multicollinearity include IC_60 and STOP_60, STOP_30, POP_15T65, HH_TOTAL, MP_HH_NC, MP_HH_WC, and CAR_OWN (see table A5.1 and table A5.2). Obviously, there is no use in including the categorical variables, since all stations stem from the same category.

Table A5.3: Casewise diagnostics of the first attempt

Case Number	Std. Residual	N2006	Predicted Value	Residual
27	8,632	21153	4197,21	16955,791

Table A5.4: Casewise diagnostics of the second attempt

Case Number	Std. Residual	N2006	Predicted Value	Residual
18	2,182	11789	9259,25	2529,748
23	2,896	12131	8774,53	3356,470
26	2,805	7724	4472,95	3251,055
45	2,387	7569	4801,95	2767,051
55	2,153	9932	7436,07	2495,928
75	-2,167	3670	6182,12	-2512,123

The first attempt yields an outlier in the case of Duiven station (see table A5.3). This case is removed before he second run, which does not show any cases with a standardized residual over ±2.9, let alone 4.0 (see table A5.4).

Table A5.5: Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
6	(Constant)	-2074,112	310,133		-6,688	,000
	FREQUENCY_oS	230,388	54,157	,267	4,254	,000
	POP_TOTAL	,115	,017	,346	6,666	,000
	BTM_TOTAL	51,653	8,891	,284	5,810	,000
	TYPE_SCORE	1204,444	270,516	,202	4,452	,000
	TOT_60	36,428	9,943	,210	3,664	,000
	TRANSFER	1027,195	509,295	,105	2,017	,046

The outcome model includes six variables, of which POP_TOTAL is the most influential (see table A5.5). All variables are significant at, at least, the 95 percent level. The coefficients make up the following function:

Y = -2074.1 + 230.4 * FREQUENCY_oS + 0.115 * POP_TOTAL + 51.7 * BTM_TOTAL + 1204.4 * TYPE_SCORE + 36.4 * TOT_60 + 1027.2 * TRANSFER

Table A5.6: ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
6	Regression	5,544E8	6	92401003,775	68,770	,000[f]
	Residual	1,559E8	116	1343630,386		
	Total	7,103E8	122			

The model is significant at the 0.000 percent level (see table A5.6). The coefficient of determination is 0.781, which means the model explains 78.1 percent of the variance in the dependent variable (see table A5.7).

Table A5.7: Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
6	,883[f]	,781	,769	1159,151

Figure A5.1: Histogram of the normal distribution of the residuals

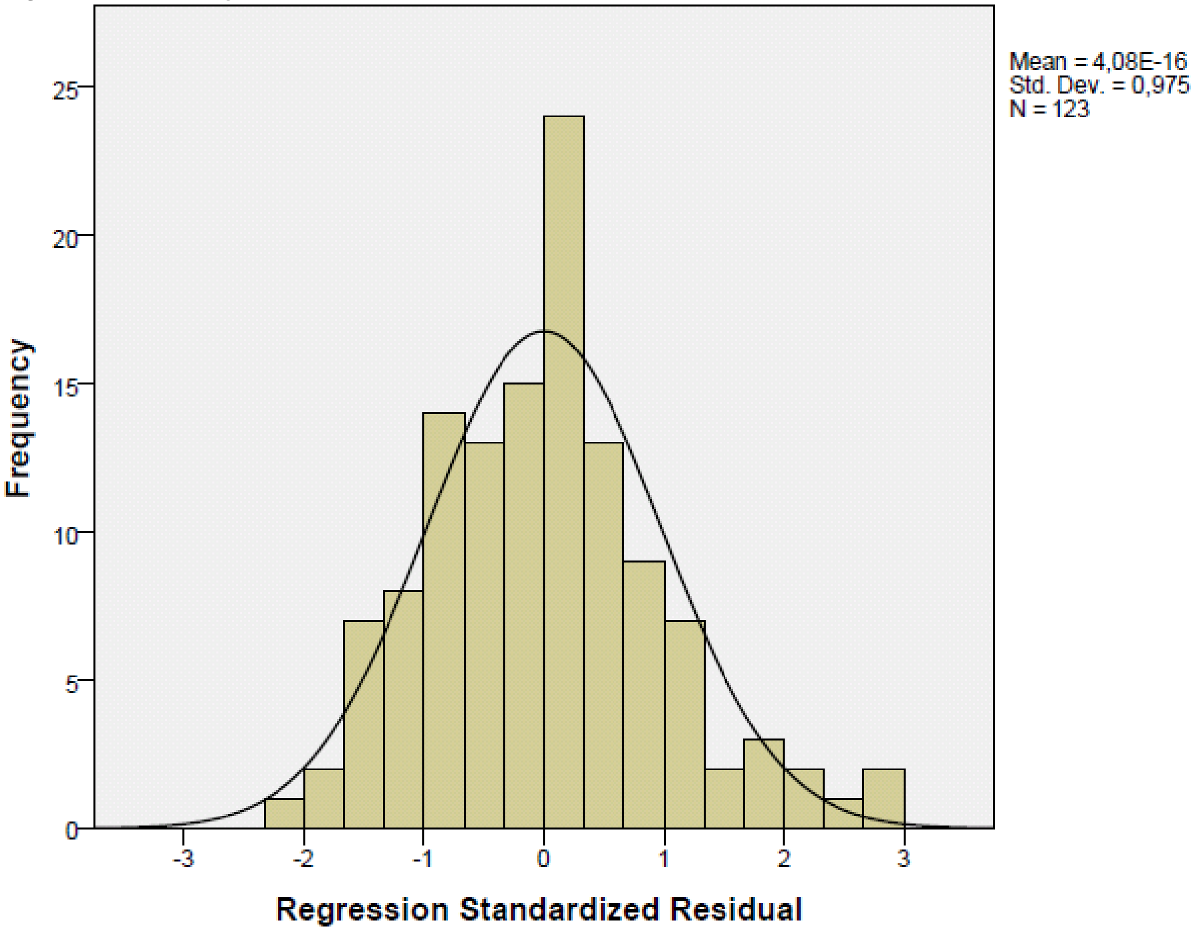


Figure A5.2: Normal probability plot of the standardized residual

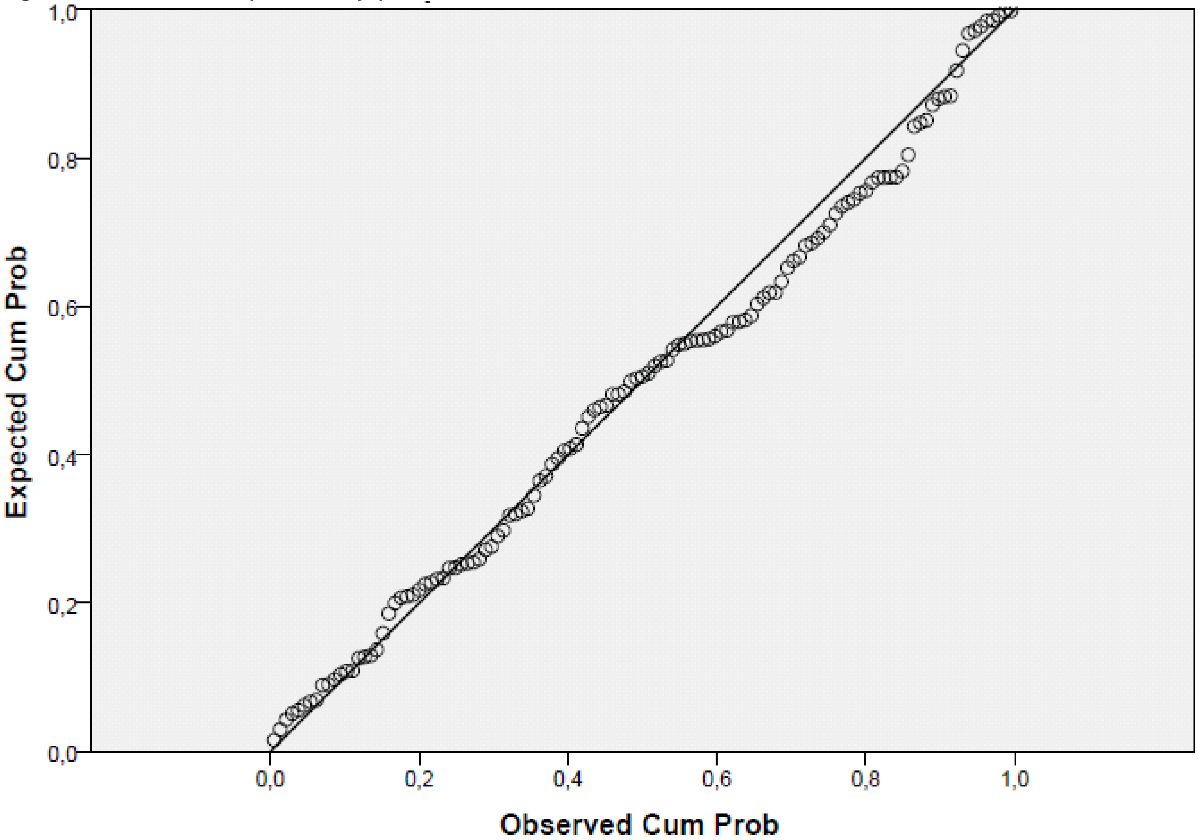
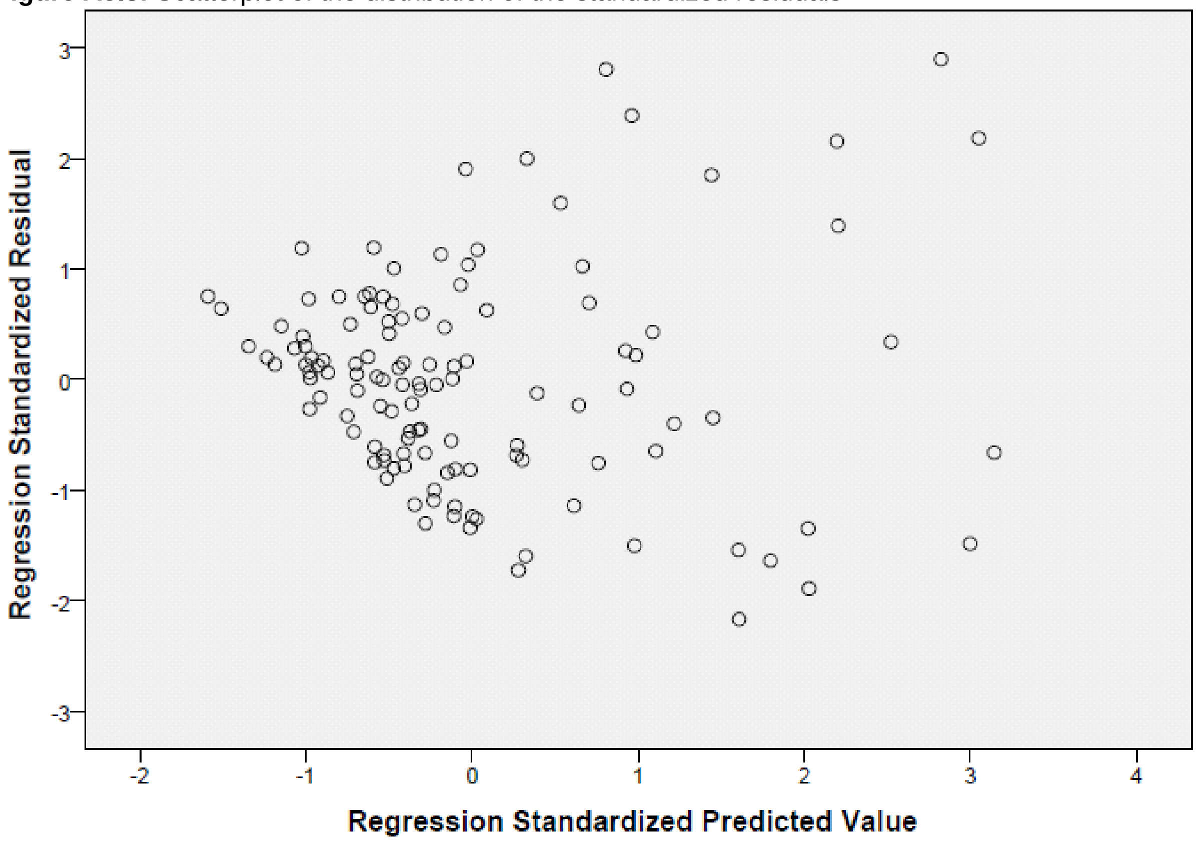


Figure A5.3: Scatterplot of the distribution of the standardized residuals



Appendix 6: SPSS output for the Category 5 stations reference class

Table A6.1: Part of the correlation matrix from SPSS

		N2006	TOT_60	IC_60	STOP_60	TOT_30
N2006	Pearson Correlation	1	,524**	,535**	,501**	,680**
	Sig. (2-tailed)		,000	,000	,000	,000
	N	87	87	87	87	87
TOT_60	Pearson Correlation	,524**	1	,939**	,990**	,796**
	Sig. (2-tailed)	,000		,000	,000	,000
	N	87	87	87	87	87
IC_60	Pearson Correlation	,535**	,939**	1	,880**	,710**
	Sig. (2-tailed)	,000	,000		,000	,000
	N	87	87	87	87	87
STOP_60	Pearson Correlation	,501**	,990**	,880**	1	,804**
	Sig. (2-tailed)	,000	,000	,000		,000
	N	87	87	87	87	87
TOT_30	Pearson Correlation	,680**	,796**	,710**	,804**	1
	Sig. (2-tailed)	,000	,000	,000	,000	
	N	87	87	87	87	87
IC_30	Pearson Correlation	,656**	,751**	,771**	,716**	,811**
	Sig. (2-tailed)	,000	,000	,000	,000	,000
	N	87	87	87	87	87
STOP_30	Pearson Correlation	,602**	,712**	,586**	,739**	,959**
	Sig. (2-tailed)	,000	,000	,000	,000	,000
	N	87	87	87	87	87

The variable TOTAL_60 is excluded due to multicollinearity because IC_60 has a higher correlation with N2006. The other variable from table A5.1 that is excluded is STOP_30. Other excluded variables include POP_15T65, HH_TOTAL, ONEP_HH, MP_HH_NC and MP_HH_WC. Obviously, there is no use in including the categorical variables, since all stations stem from the same category.

Table A6.2: Casewise diagnostics of the first model run

Case Number	Std. Residual	N2006	Predicted Value	Residual
24	3,107	12465	4817,81	7647,189
30	-3,685	4126	13197,97	-9071,973
52	4,290	39555	28995,64	10559,364
61	-2,545	1906	8170,80	-6264,802

As shown by table A6.2, the first model run gave one case with a standardized residual of more than 4. This case is Amsterdam Sloterdijk. This is case subsequently removed. The second and third run both also lead to the removal of a single case (see table A6.3 and table A6.4). These are Schiedam Centrum and Rotterdam Blaak. The fourth run does not yield any outliers (see table A6.5).

Table A6.3: Casewise diagnostics of the second model run

Case Number	Std. Residual	N2006	Predicted Value	Residual
3	3,477	12574	6465,89	6108,107
24	4,568	12465	4439,47	8025,526
28	2,391	6380	2178,81	4201,193
30	-2,390	4126	8324,67	-4198,667

Table A6.4: Casewise diagnostics of the third model run

Case Number	Std. Residual	N2006	Predicted Value	Residual
3	5,146	12574	5318,70	7255,297
24	2,091	4753	1804,60	2948,400
27	2,362	6380	3049,19	3330,807

Table A6.5: Casewise diagnostics of the fourth model run

Case Number	Std. Residual	N2006	Predicted Value	Residual
5	-2,175	1317	3806,20	-2489,203
23	2,696	4753	1668,52	3084,480
26	3,033	6380	2908,89	3471,105
45	2,564	10827	7893,32	2933,682

After the fourth run, the model is accepted.

Table A6.6: Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
4	(Constant)	-1130,117	534,502		-2,114	,038
	IC_30	325,466	129,964	,269	2,504	,014
	BTM_TOTAL	28,718	6,557	,368	4,380	,000
	IC_60	95,017	34,081	,255	2,788	,007
	BUILTUP_P	24,079	9,353	,164	2,574	,012

The coefficients of table A6.6 lead to the following function for the reference class of category 5 stations:

Y = -1130.1 + 325.5 * IC_30 + 28.7 * BTM_TOTAL + 95.0 * IC_60 + 24.1 * BUILTUP_P

The model includes four variables, of which BTM_TOTAL is the most influential. All variable are significant at, at least, the 95 percent level.

The model is significant at the 0.000 percent level (see table A6.7). It has an R^2 of 0.722, which means 72.2 percent of the variance of the dependent variable is explained by the model (see table A6.8).

Table A6.7: ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
4	Regression	2,689E8	4	67228423,460	51,344	,000[d]
	Residual	1,034E8	79	1309385,192		
	Total	3,724E8	83			

Table A6.8: Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,772[a]	,596	,591	1354,135
2	,824[b]	,679	,671	1215,005
3	,836[c]	,699	,688	1183,845
4	,850[d]	,722	,708	1144,284

Figure A6.1: Histogram of the normal distribution of the residuals

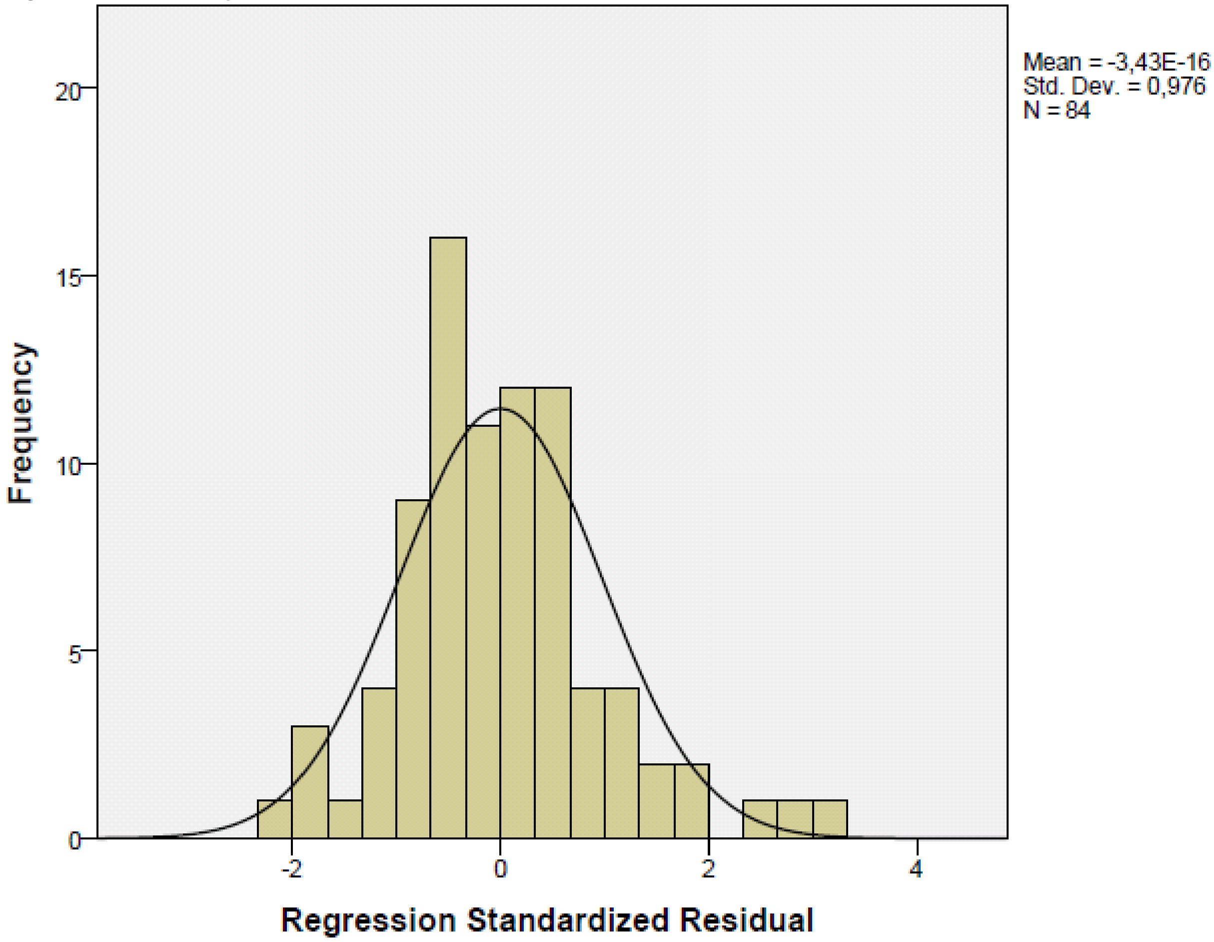


Figure A6.2: Normal probability plot of the standardized residual

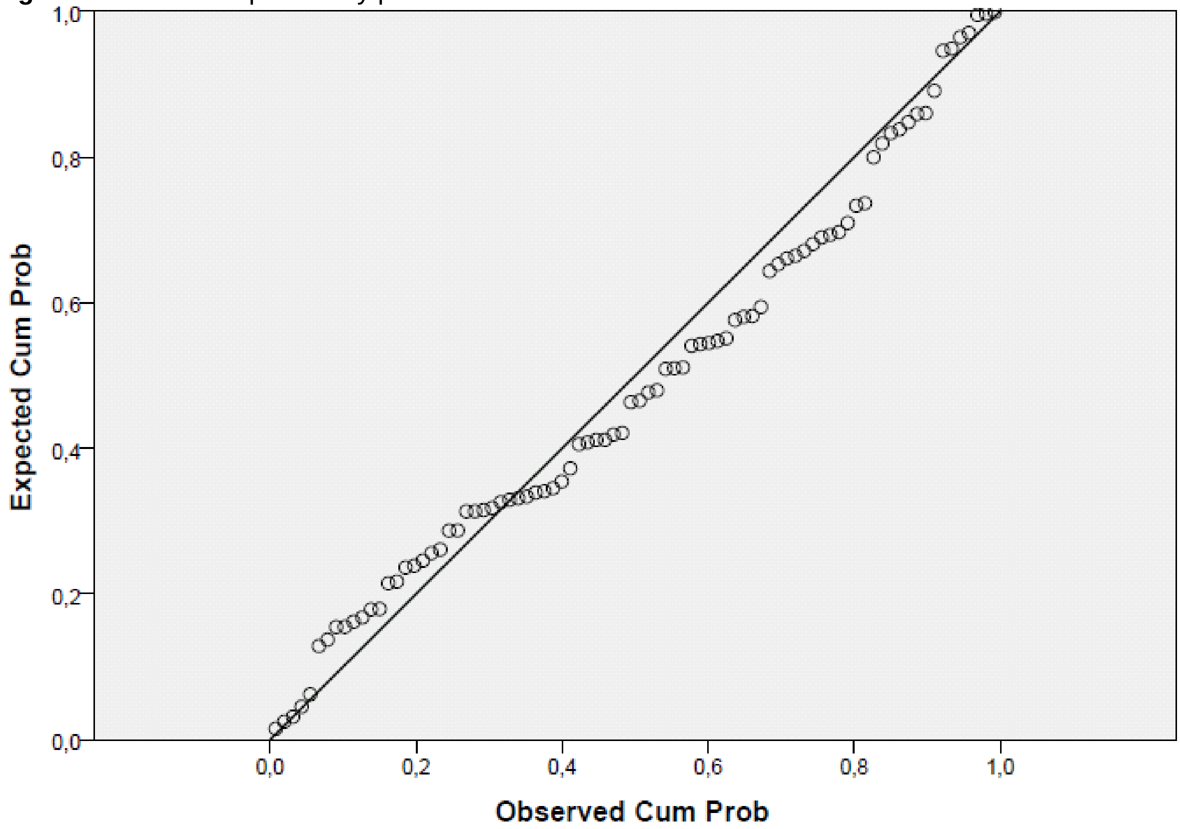
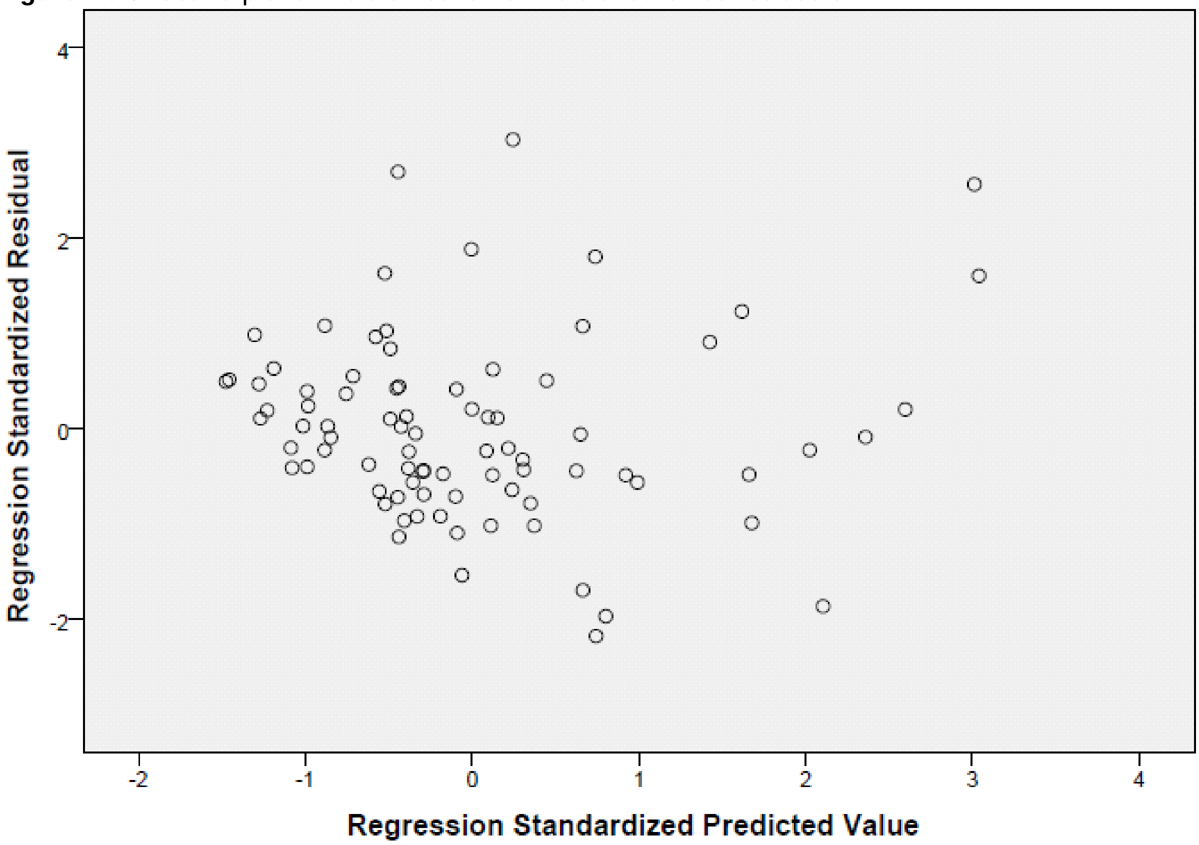


Figure A6.3: Scatterplot of the distribution of the standardized residuals



Appendix 7: SPSS output for the Category 6 stations reference class

Table A7.1: Part of the correlation matrix from SPSS

		N2006	TOT_60	IC_60	STOP_60	TOT_30
STOP_60	Pearson Correlation	,404**	,975**	,730**	1	,643**
	Sig. (2-tailed)	,000	,000	,000		,000
	N	85	85	85	85	85

Table A7.2: Part of the correlation matrix from SPSS

		CII_RES_RATE	POP_TOTAL	POP_15T65
POP_15T65	Pearson Correlation	-,097	,999**	1
	Sig. (2-tailed)	,377	,000	
	N	85	85	85
POP_ALLOCH	Pearson Correlation	,038	,842**	,843**
	Sig. (2-tailed)	,728	,000	,000
	N	85	85	85
HH_TOTAL	Pearson Correlation	-,091	,994**	,993**
	Sig. (2-tailed)	,408	,000	,000
	N	85	85	85
ONEP_HH	Pearson Correlation	-,056	,957**	,954**
	Sig. (2-tailed)	,612	,000	,000
	N	85	85	85
MP_HH_NC	Pearson Correlation	-,103	,986**	,985**
	Sig. (2-tailed)	,350	,000	,000
	N	85	85	85
MP_HH_WC	Pearson Correlation	-,108	,995**	,996**
	Sig. (2-tailed)	,327	,000	,000
	N	85	85	85
CAR_OWN	Pearson Correlation	-,102	,994**	,993**
	Sig. (2-tailed)	,355	,000	,000
	N	85	85	85

Variables excluded due to multicollinearity include STOP_60 (see table A7.1), POP_15T65, HH_TOTAL, ONEP_HH, MP_HH_NC, MP_HH_WC, and CAR_OWN. Obviously, there is no use in including the categorical variables, since all stations stem from the same category.

The first run of the model does not show any cases with a standardized residual of more than 4 (see table A7.3). The case with the highest standardized residual is Boxtel, with a predicted value which is 2050 lower than the actual ridership of 2006.

The model is accepted.

Table A7.3: Casewise diagnostics

Case Number	Std. Residual	N2006	Predicted Value	Residual
9	3,744	5486	3435,69	2050,311
11	-2,031	410	1522,09	-1112,091
15	2,422	3010	1683,51	1326,491
67	-2,089	313	1457,13	-1144,128

Table A7.4: Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
5	(Constant)	-877,520	198,645		-4,418	,000
	POP_TOTAL	,130	,015	,523	8,889	,000
	FREQUENCY_oS	317,779	44,647	,490	7,118	,000
	STOP_30	-109,092	34,036	-,200	-3,205	,002
	BTM_TOTAL	52,634	17,503	,184	3,007	,004
	URBANIZATION_3	473,572	220,150	,123	2,151	,035

The function that can be derived from the coefficients from table A7.4 is as follows:

Y = -877.5 + 0.130 * POP_TOTAL + 317.8 * FREQUENCY_oS -109.1 * STOP_30 + 52.6 * BTM_TOTAL + 473.6 * URBANIZATION_3

All variables are significant at, at least, the 95 percent level. The model is significant at the 0.000 percent level (see table A7.5). The model has an R^2 of 0.786, which means it explains 78.6 percent of the variance in the dependent variable N2006 (see table A7.6).

Table A7.5: ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
5	Regression	83547730,940	5	16709546,188	55,724	,000[e]
	Residual	22789462,572	76	299861,350		
	Total	1,063E8	81			

Table A7.6: Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,743[a]	,552	,546	772,032
2	,857[b]	,735	,728	597,528
3	,869[c]	,756	,746	577,178
4	,879[d]	,773	,761	560,346
5	,886[e]	,786	,772	547,596

Figure A7.1: Histogram of the normal distribution of the residuals

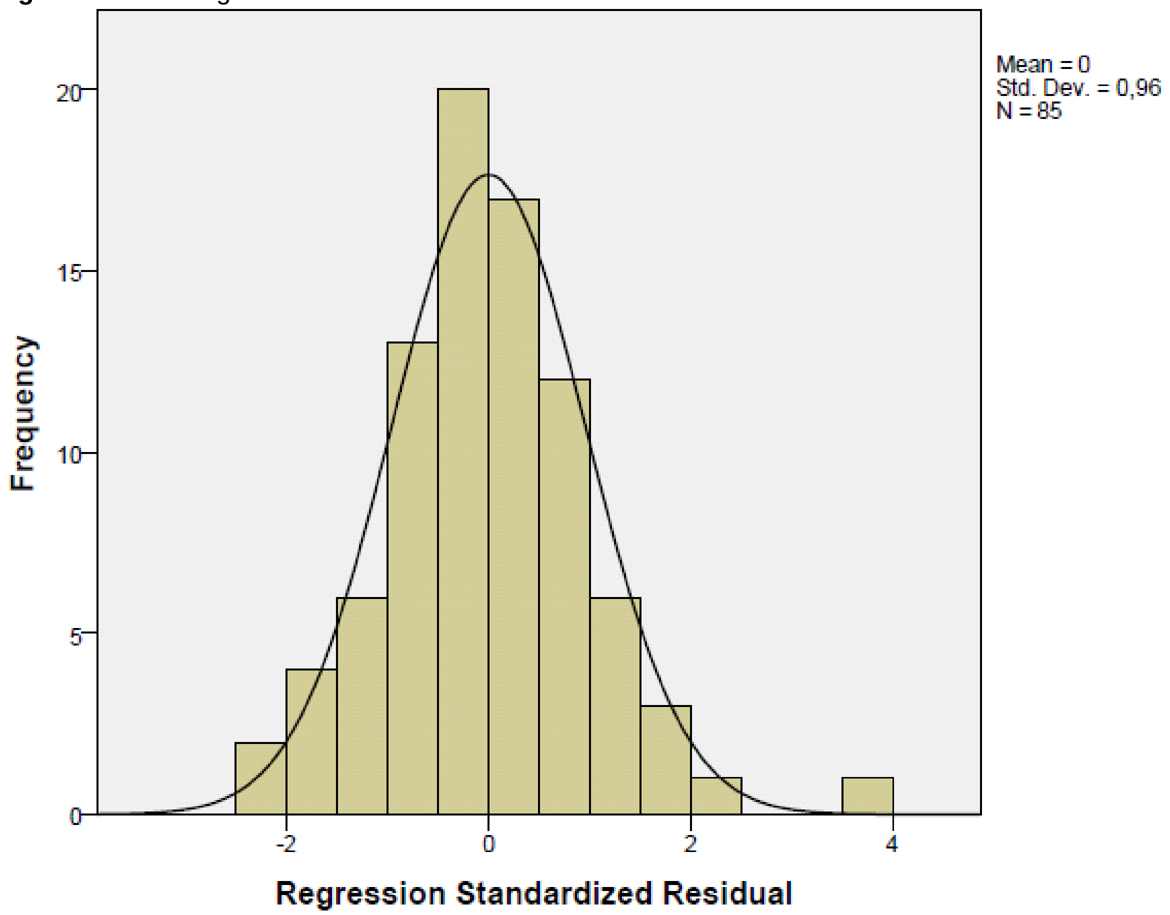


Figure A7.2: Normal probability plot of the standardized residual

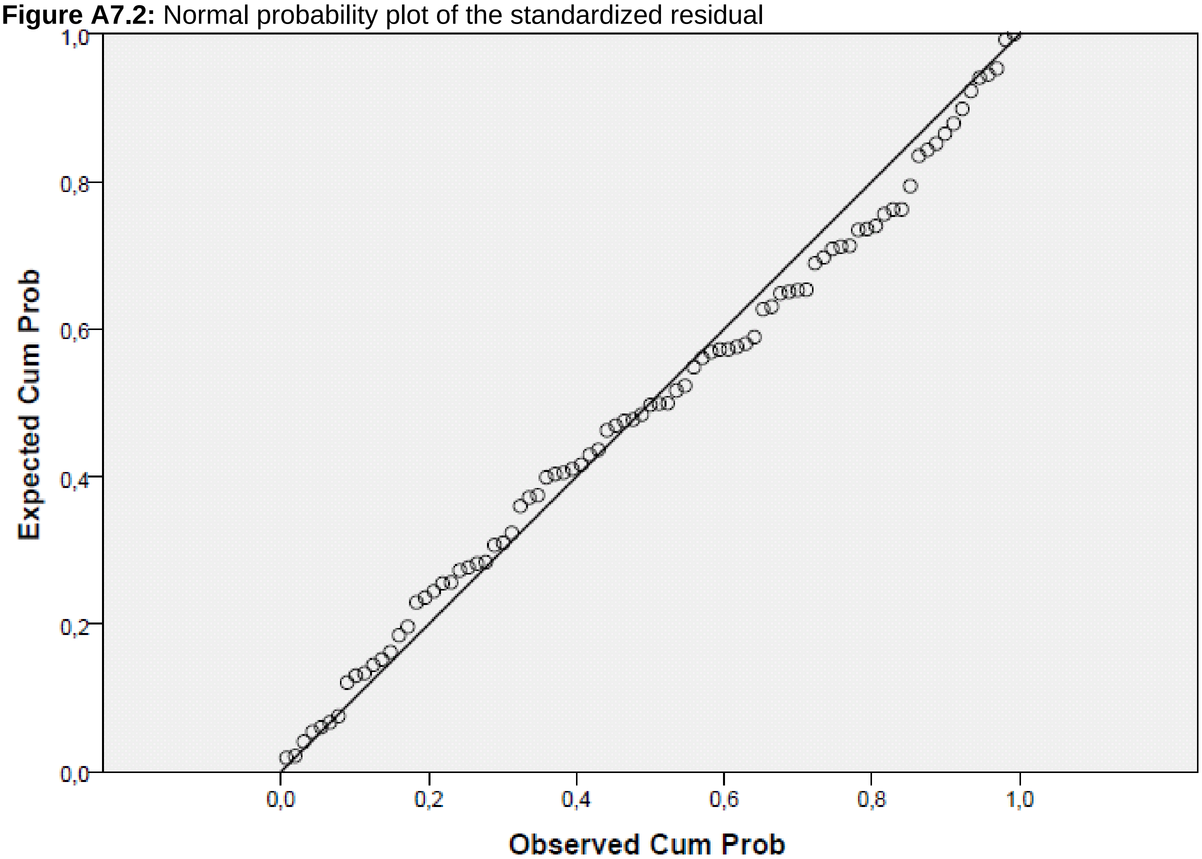
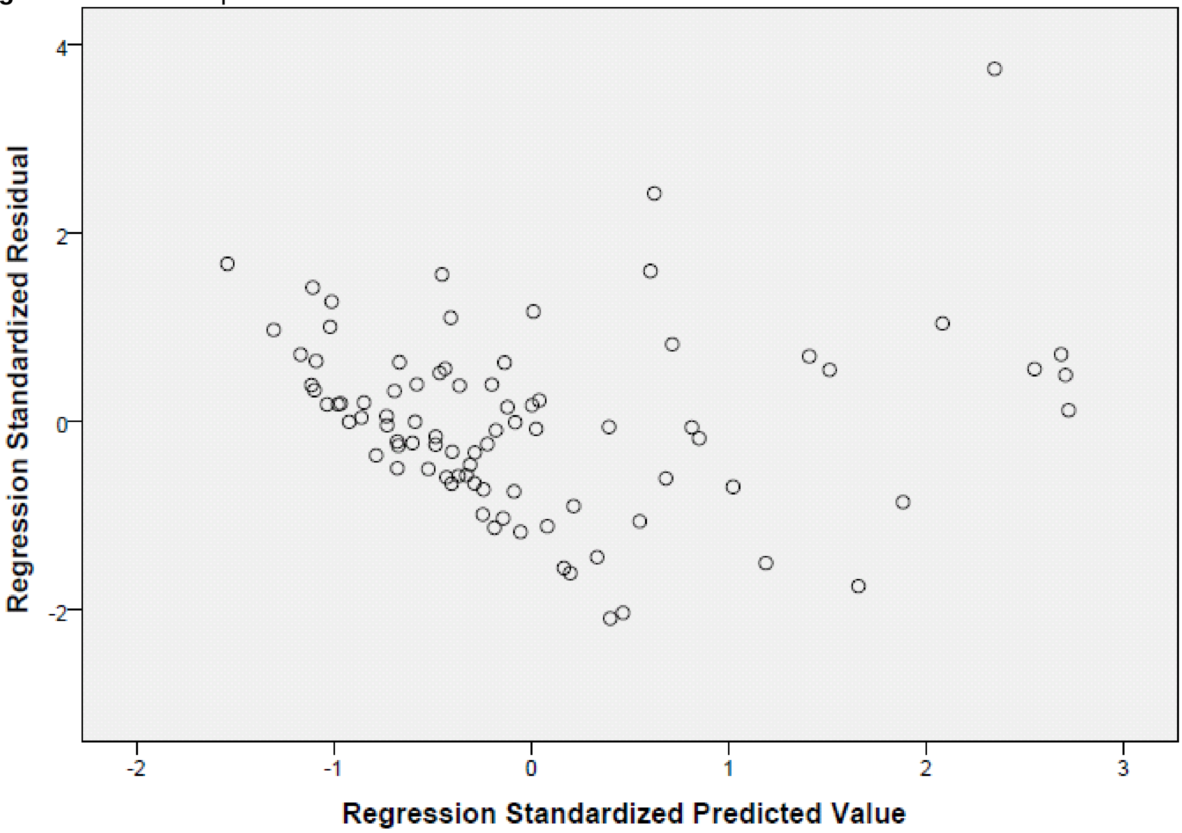


Figure A7.3: Scatterplot of the distribution of the standardized residuals



Appendix 8: SPSS output for the Non Randstad Category 4 stations reference class

Table A8.1: Part of the correlation matrix from SPSS

		N2006	TOT_60	IC_60	STOP_60	TOT_30
N2006	Pearson Correlation	1	,296**	,342**	,235*	,211*
	Sig. (2-tailed)		,004	,001	,022	,041
	N	94	94	94	94	94
TOT_60	Pearson Correlation	,296**	1	,774**	,961**	,702**
	Sig. (2-tailed)	,004		,000	,000	,000
	N	94	94	94	94	94
IC_60	Pearson Correlation	,342**	,774**	1	,569**	,467**
	Sig. (2-tailed)	,001	,000		,000	,000
	N	94	94	94	94	94
STOP_60	Pearson Correlation	,235*	,961**	,569**	1	,708**
	Sig. (2-tailed)	,022	,000	,000		,000
	N	94	94	94	94	94
TOT_30	Pearson Correlation	,211*	,702**	,467**	,708**	1
	Sig. (2-tailed)	,041	,000	,000	,000	
	N	94	94	94	94	94
IC_30	Pearson Correlation	,030	,514**	,596**	,408**	,410**
	Sig. (2-tailed)	,775	,000	,000	,000	,000
	N	94	94	94	94	94
STOP_30	Pearson Correlation	,219*	,560**	,269**	,610**	,929**
	Sig. (2-tailed)	,034	,000	,009	,000	,000
	N	94	94	94	94	94
CROSSINGS	Pearson Correlation	,419**	,100	,245*	,023	-,023
	Sig. (2-tailed)	,000	,338	,017	,824	,828
	N	94	94	94	94	94
BUILTUP_P	Pearson Correlation	,372**	,029	,122	-,016	-,010
	Sig. (2-tailed)	,000	,782	,241	,881	,924
	N	94	94	94	94	94

Table A8.2: Part of the correlation matrix from SPSS

		POP_ALLOCH	HH_TOTAL	ONEP_HH	MP_HH_NC
POP_TOTAL	Pearson Correlation	,801**	,985**	,875**	,977**
	Sig. (2-tailed)	,000	,000	,000	,000
	N	94	94	94	94
POP_15T65	Pearson Correlation	,813**	,982**	,868**	,974**
	Sig. (2-tailed)	,000	,000	,000	,000
	N	94	94	94	94

STOP_60 and TOT_30 are among the variables excluded due to milticollinearity (see table A8.1). Others include CROSSINGS, POP_15T65, HH_TOTAL, MP_HH_NC, MP_HH_WC, and CAR_OWN due to multicollinearity with POP_TOTAL (see table A8.2 and table A8.3).

Because all stations in this reference class are of category 4 and are located outside of the Randstad area, the variables RANDSTAD, CATEGORY_5, and CATEGORY_6 can be left out of the MR analysis. These simply do not diversify. For the same reason, URBANIZATION_1 can be disregarded. None of the stations in this reference class is located within a municipality with the highest degree of urbanization.

Table A8.3: Part of the correlation matrix from SPSS

		CII_RES_RATE	POP_TOTAL	POP_15T65
POP_ALLOCH	Pearson Correlation	-,138	,801**	,813**
	Sig. (2-tailed)	,183	,000	,000
	N	94	94	94
HH_TOTAL	Pearson Correlation	-,242*	,985**	,982**
	Sig. (2-tailed)	,019	,000	,000
	N	94	94	94
ONEP_HH	Pearson Correlation	-,146	,875**	,868**
	Sig. (2-tailed)	,159	,000	,000
	N	94	94	94
MP_HH_NC	Pearson Correlation	-,259*	,977**	,974**
	Sig. (2-tailed)	,012	,000	,000
	N	94	94	94
MP_HH_WC	Pearson Correlation	-,292**	,988**	,990**
	Sig. (2-tailed)	,004	,000	,000
	N	94	94	94
CAR_OWN	Pearson Correlation	-,276**	,983**	,976**
	Sig. (2-tailed)	,007	,000	,000
	N	94	94	94

Table A8.4: Casewise diagnostics of the first run

Case Number	Std. Residual	N2006	Predicted Value	Residual
17	8,113	21153	4264,55	16888,446

Table A8.5: Casewise diagnostics of the second run

Case Number	Std. Residual	N2006	Predicted Value	Residual
10	-2,114	1592	3568,69	-1976,692
16	3,118	7724	4808,97	2915,031
28	4,250	7569	3595,33	3973,669
30	2,273	5771	3645,68	2125,318

It is hardly surprising that Duiven station turns out to be an extreme outlier within this reference class during the first run (see table A8.4). The second run shows Heerhugowaard to be outside of the standardized residual threshold of 4 (see table A8.5).

Table A8.6: Casewise diagnostics of the third run

Case Number	Std. Residual	N2006	Predicted Value	Residual
10	-2,560	1592	3623,68	-2031,679
16	2,947	7724	5384,82	2339,175
26	2,254	4876	3087,02	1788,981
29	2,753	5771	3586,27	2184,731

The population used for the third run does not contain outliers (see table A8.6).

Table A8.7: Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
9	(Constant)	-890,581	252,135		-3,532	,001
	FREQUENCY_oS	152,303	45,110	,209	3,376	,001
	BTM_TOTAL	57,523	8,754	,383	6,571	,000
	IC_60	212,469	40,453	,397	5,252	,000
	IC_30	-503,459	115,330	-,297	-4,365	,000
	TYPE_SCORE	682,612	209,667	,169	3,256	,002
	POP_TOTAL	,118	,014	,499	8,464	,000
	JOBS_MUN	-,016	,007	-,146	-2,488	,015

The coefficients resulting from the final run, which are presented in table A8.7, lead to the following function:

Y = -890.6 + 152.3 * FREQUENCY_oS + 57.5 * BTM_TOTAL + 212.5 * IC_60 -503.5 * IC_30 + 682.6 * TYPE_SCORE + 0.118 * POP_TOTAL -0.016 * JOBS_MUN

The model contains seven variables, of which POP_TOTAL has the most influence on the dependent variable. All variables are significant at, at least, the 95 percent level (see table A8.7). The model is significant at the 0.000 percent level and explains 78.4 percent of the variance in the dependent variable (see table A8.8).

Table A8.8: ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
9	Regression	1,926E8	7	27514297,000	43,679	,000[i]
	Residual	52912954,078	84	629916,120		
	Total	2,455E8	91			

Table A8.8: Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
9	,886[i]	,784	,767	793,673

Figure A8.1: Histogram of the normal distribution of the residuals

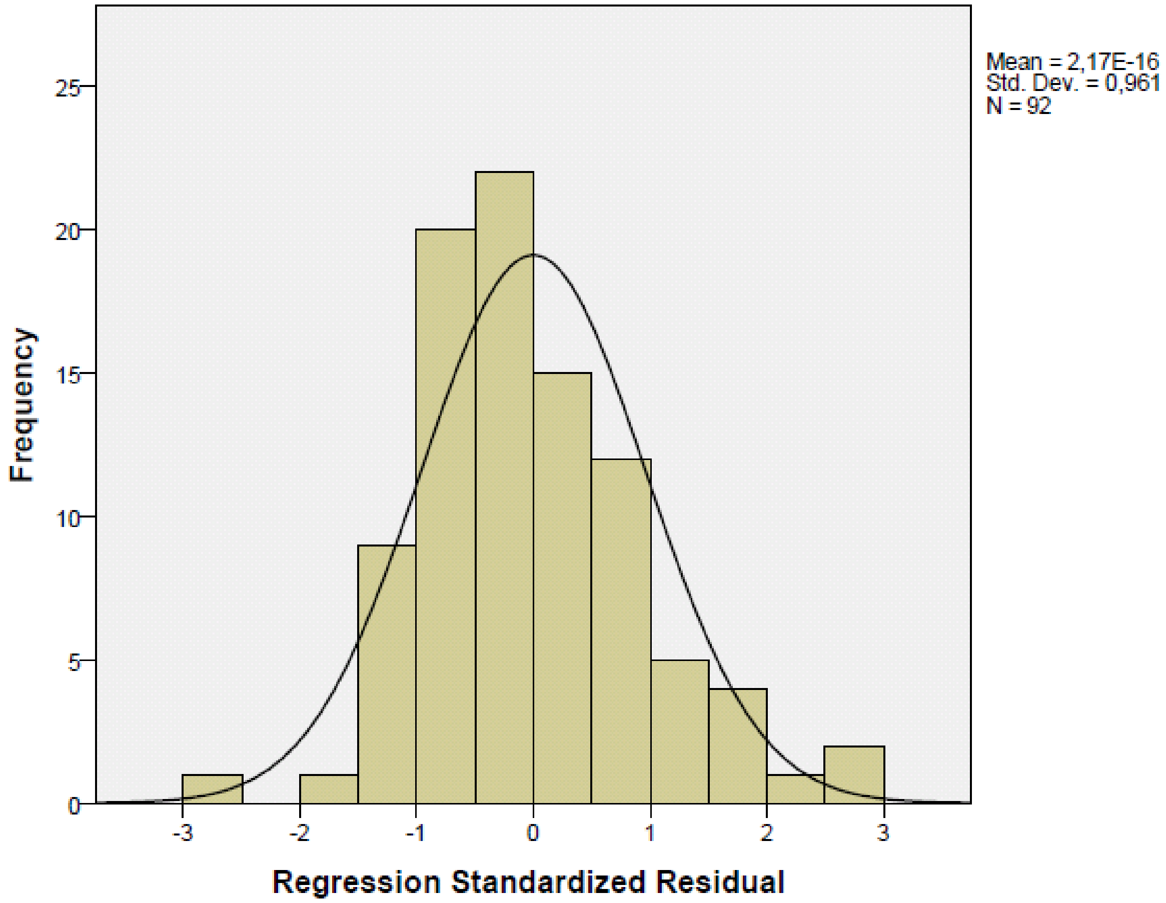


Figure A8.2: Normal probability plot of the standardized residual

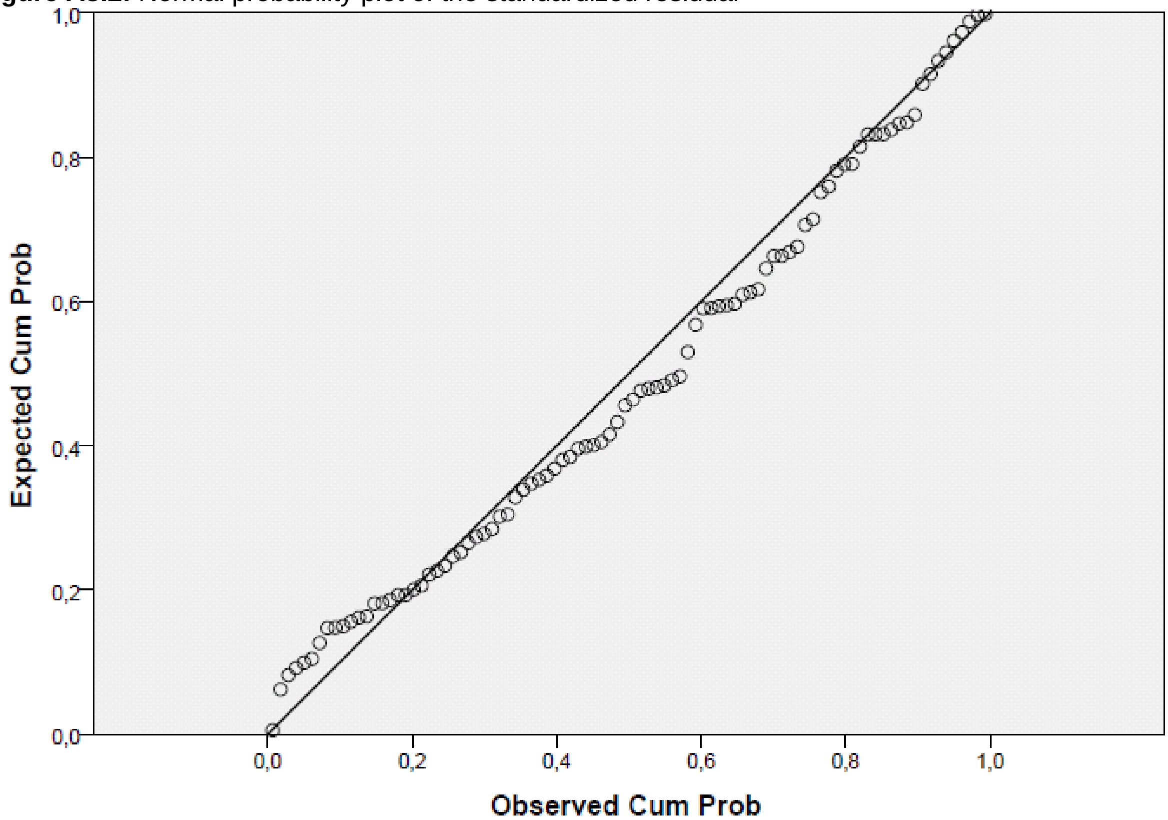


Figure A8.3: Scatterplot of the distribution of the standardized residuals

Appendix 9: SPSS output for the Non Randstad Category 5 stations reference class

Table A9.1: Part of the correlation matrix from SPSS

		N2006	TOT_60	IC_60	STOP_60	TOT_30
N2006	Pearson Correlation	1	,515**	,548**	,455**	,449**
	Sig. (2-tailed)		,000	,000	,002	,002
	N	45	45	45	45	45
TOT_60	Pearson Correlation	,515**	1	,844**	,975**	,727**
	Sig. (2-tailed)	,000		,000	,000	,000
	N	45	45	45	45	45
IC_60	Pearson Correlation	,548**	,844**	1	,704**	,451**
	Sig. (2-tailed)	,000	,000		,000	,002
	N	45	45	45	45	45
STOP_60	Pearson Correlation	,455**	,975**	,704**	1	,776**
	Sig. (2-tailed)	,002	,000	,000		,000
	N	45	45	45	45	45
TOT_30	Pearson Correlation	,449**	,727**	,451**	,776**	1
	Sig. (2-tailed)	,002	,000	,002	,000	
	N	45	45	45	45	45
IC_30	Pearson Correlation	,472**	,555**	,628**	,476**	,428**
	Sig. (2-tailed)	,001	,000	,000	,001	,003
	N	45	45	45	45	45
STOP_30	Pearson Correlation	,308*	,582**	,248	,667**	,933**
	Sig. (2-tailed)	,040	,000	,100	,000	,000
	N	45	45	45	45	45
CROSSINGS	Pearson Correlation	,565**	,277	,279	,251	,312*
	Sig. (2-tailed)	,000	,066	,063	,097	,037
	N	45	45	45	45	45
BUILTUP_P	Pearson Correlation	,517**	,467**	,441**	,436**	,517**
	Sig. (2-tailed)	,000	,001	,002	,003	,000
	N	45	45	45	45	45

Table A9.2: Part of the correlation matrix from SPSS

		CII_RES_RATE	POP_TOTAL	POP_15T65
POP_TOTAL	Pearson Correlation	-,327*	1	,993**
	Sig. (2-tailed)	,028		,000
	N	45	45	45
POP_15T65	Pearson Correlation	-,304*	,993**	1
	Sig. (2-tailed)	,042	,000	

TOT_60, TOP_30, POP_15T65, HH_TOTAL, MP_HH_NC, and MP_HH_WC are excluded due to multicollinearity (see table A9.1 through table A9.3).

Table A9.3: Part of the correlation matrix from SPSS

		CII_RES_RATE	POP_TOTAL	POP_15T65
POP_15T65	N	45	45	45
POP_ALLOCH	Pearson Correlation	-,202	,856**	,850**
	Sig. (2-tailed)	,183	,000	,000
	N	45	45	45
HH_TOTAL	Pearson Correlation	-,263	,961**	,977**
	Sig. (2-tailed)	,081	,000	,000
	N	45	45	45
ONEP_HH	Pearson Correlation	-,162	,829**	,870**
	Sig. (2-tailed)	,288	,000	,000
	N	45	45	45
MP_HH_NC	Pearson Correlation	-,333*	,982**	,973**
	Sig. (2-tailed)	,025	,000	,000
	N	45	45	45
MP_HH_WC	Pearson Correlation	-,364*	,902**	,869**
	Sig. (2-tailed)	,014	,000	,000
	N	45	45	45
CAR_OWN	Pearson Correlation	-,306*	,826**	,805**
	Sig. (2-tailed)	,041	,000	,000
	N	45	45	45

Table A9.4: Casewise diagnostics

Case Number	Std. Residual	N2006	Predicted Value	Residual
6	-2,380	1159	3223,24	-2064,236
8	2,267	4753	2787,08	1965,916
11	3,776	6380	3105,78	3274,223

Table A9.5: Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
3	(Constant)	-1519,087	423,069		-3,591	,001
	FREQUENCY_oS	242,580	64,791	,426	3,744	,001
	CROSSINGS	2,380	,652	,375	3,648	,001
	IC_60	89,601	43,987	,231	2,037	,048

There are no cases with a standardized residual over 4 within this reference class (see table A9.4). The highest standardized residual belongs to the station Tilburg Universiteit, formerly known as Tilburg West. Nevertheless, the model is accepted and the coefficients from table A9.5 lead to the following function:

Y = -1519.1 + 242.6 * FREQUENCY_oS + 2.38 * CROSSINGS + 89.6 * IC_60

The model consists of three variables which are all within the 95 percent level (see table A9.5). Table A9.6 shows that the model is significant at the 0.000 percent level. The model accounts for 61.5 percent of the variance in the dependent variable (see table A9.7).

Table A9.6: ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
3	Regression	49354282,214	3	16451427,405	21,876	,000[c]
	Residual	30833877,563	41	752045,794		
	Total	80188159,778	44			

Table A9.7: Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,652[a]	,425	,411	1035,775
2	,759[b]	,577	,556	899,131
3	,785[c]	,615	,587	867,206

Figure A9.1: Histogram of the normal distribution of the residuals

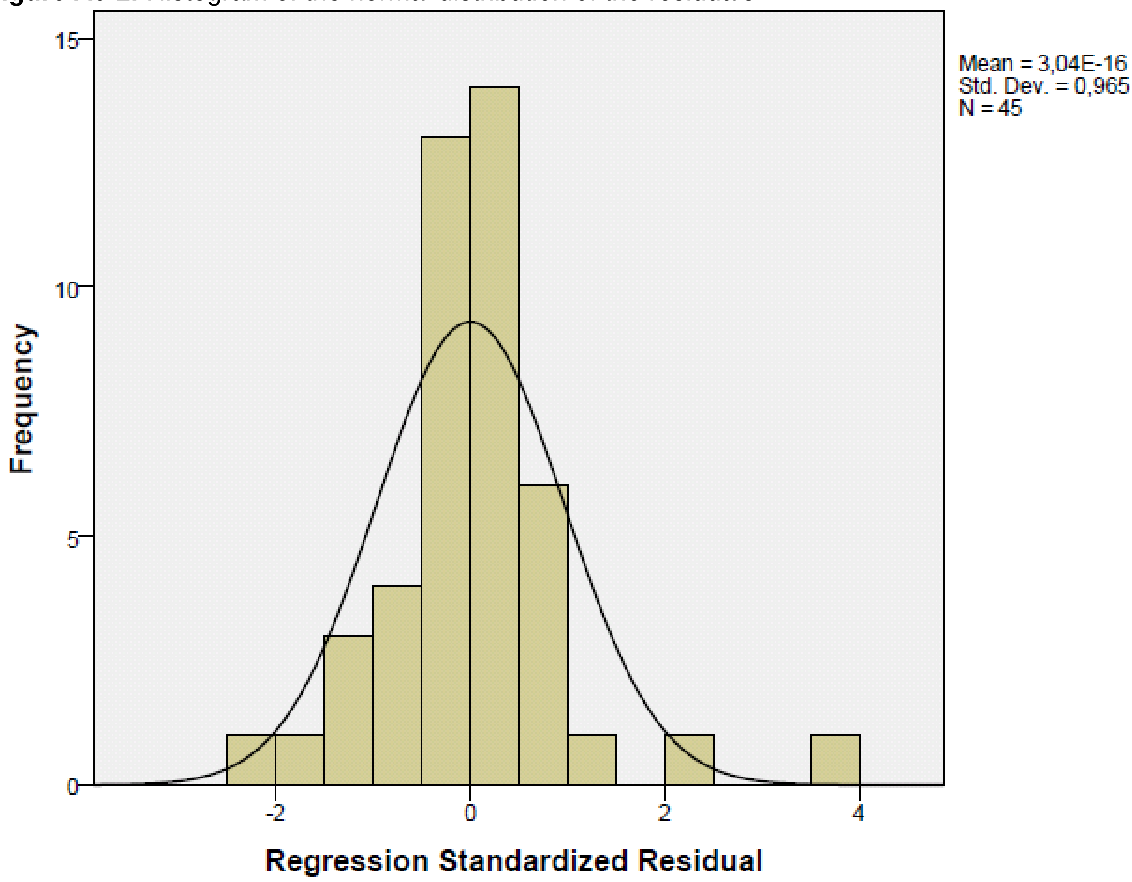


Figure A9.2: Normal probability plot of the standardized residual

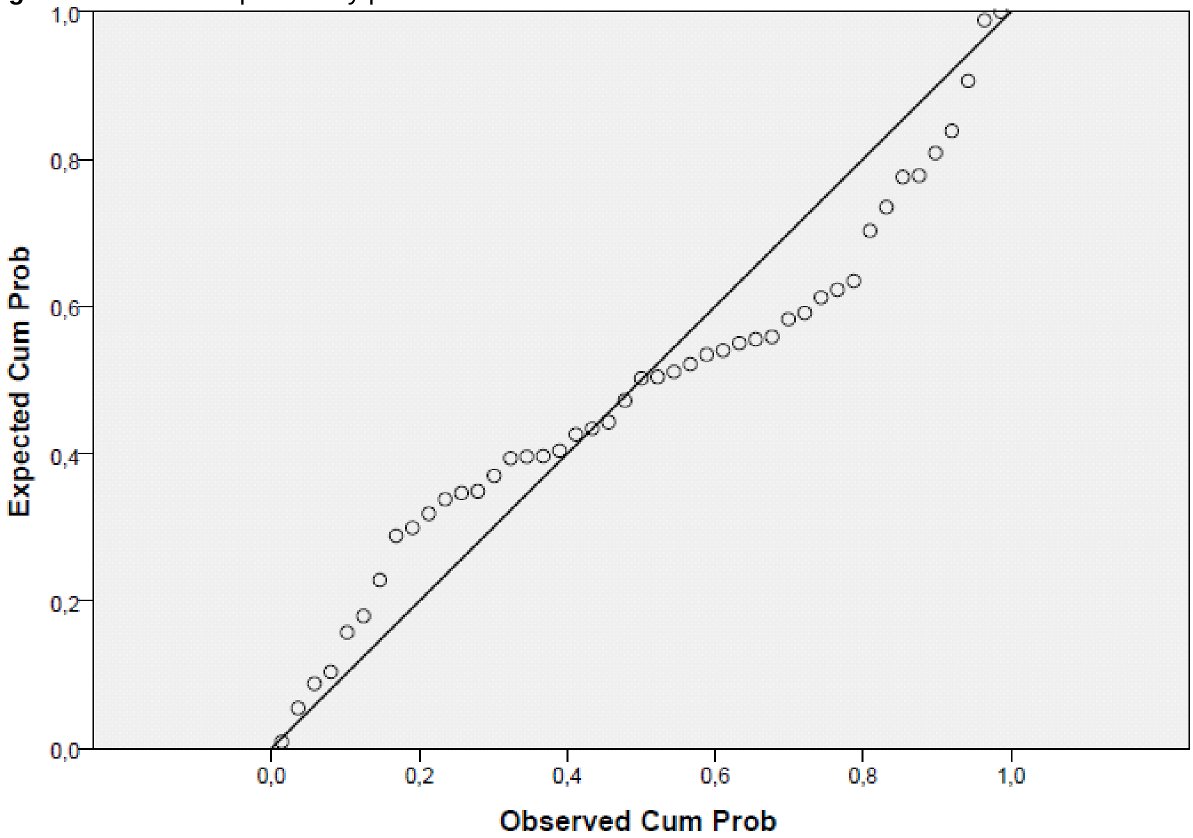
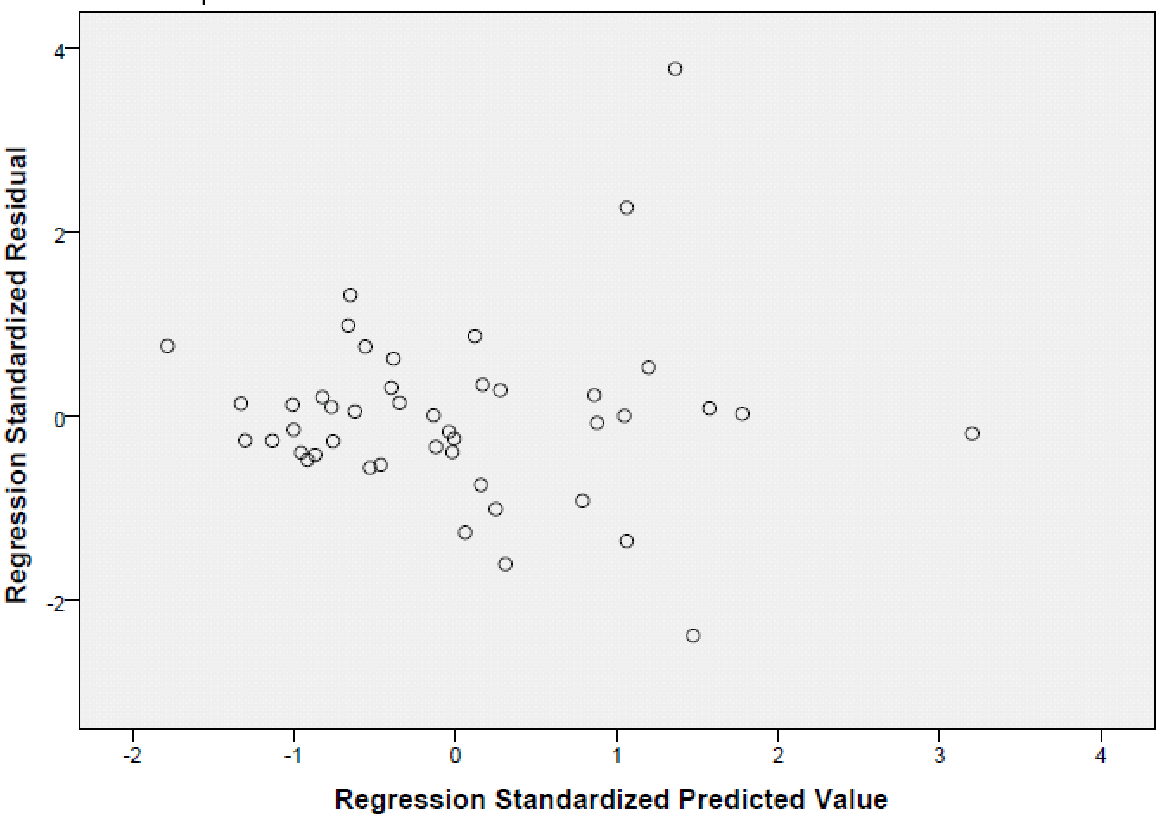


Figure A9.3: Scatterplot of the distribution of the standardized residuals



Though the model can be accepted because there are no cases with a standardized residual over 4, the model does leave room for improvement, judging by the relatively low R^2. It makes sense to add a variable for the university near the station Tilburg Universiteit, especially since this is the station with the highest standardized residual. The outputs resulting from adding this dichotomous variable, as UNIVERSITY, are presented below.

Table A9.8: Casewise diagnostics after adding the variable UNIVERSITY

Case Number	Std. Residual	N2006	Predicted Value	Residual
6	-2,706	1159	3037,86	-1878,858
8	2,813	4753	2799,39	1953,606
27	-2,006	599	1991,91	-1392,913

Table A9.9: Coefficients after adding the variable UNIVERSITY

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
4	(Constant)	-5231,915	1812,945		-2,886	,006
	FREQUENCY_oS	276,708	46,454	,486	5,957	,000
	UNIVERSITY	3796,114	722,244	,419	5,256	,000
	CROSSINGS	1,756	,558	,276	3,150	,003
	INCOME_AV	,335	,148	,189	2,259	,029

Y = -5231.9 + 276.7 * FREQUENCY_oS + 3796.1 * UNIVERSITY + 1.756 * CROSSINGS + 0.335 * INCOME_AV

Table A9.9: ANOVA after adding the variable UNIVERSITY

Model		Sum of Squares	df	Mean Square	F	Sig.
4	Regression	60901080,789	4	15225270,197	31,576	,000[d]
	Residual	19287078,989	40	482176,975		
	Total	80188159,778	44			

Table A9.10: Model summary after adding the variable UNIVERSITY

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,652[a]	,425	,411	1035,775
2	,788[b]	,622	,604	850,014
3	,854[c]	,729	,709	728,310
4	,871[d]	,759	,735	694,390

Adding the variable UNIVERSITY leads to a higher coefficient of determination, a lower standard error (see table A9.10) and a lower mean square (see table A9.7). Furthermore, the normal probability plot of the standardized residual is divided more closely along the diagonal (see figure A9.5).

Figure A9.4: Histogram of the normal distribution of the residuals after adding the variable UNIVERSITY

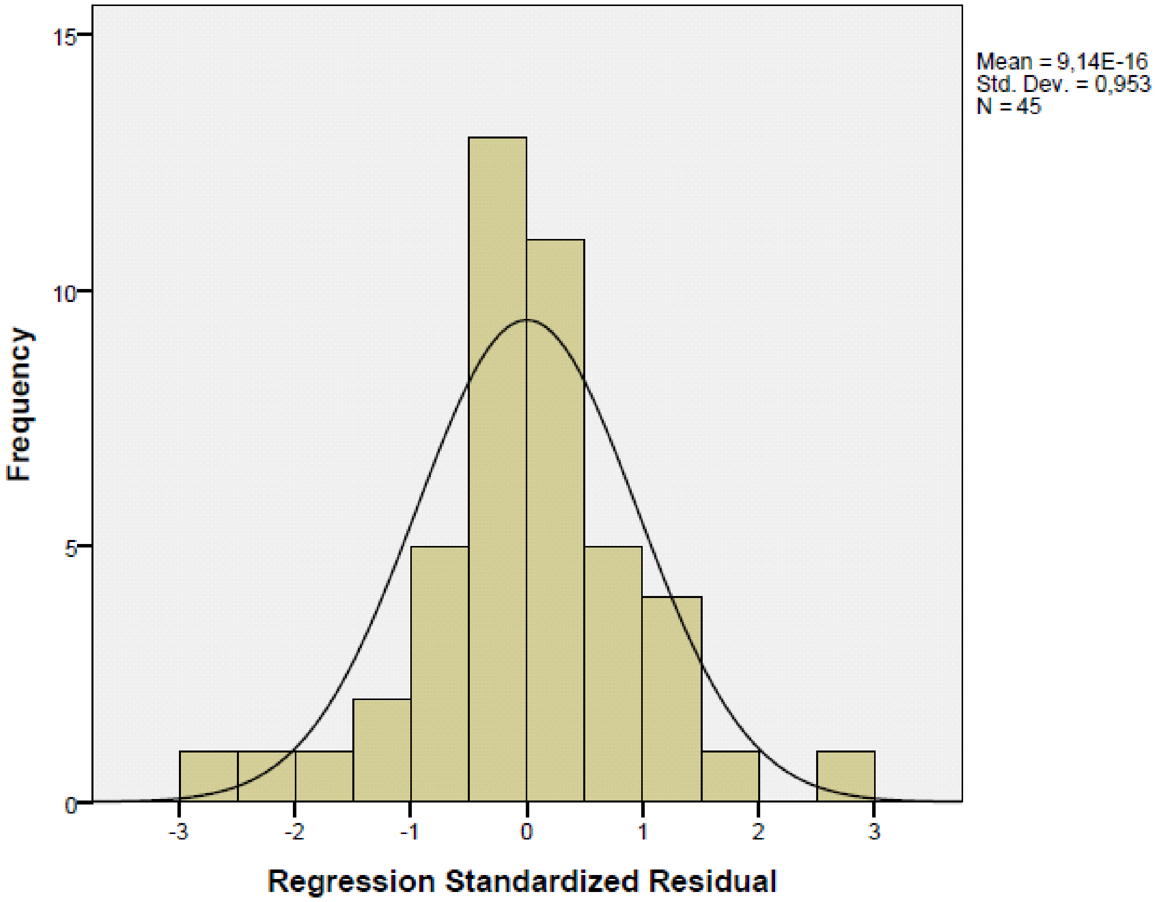
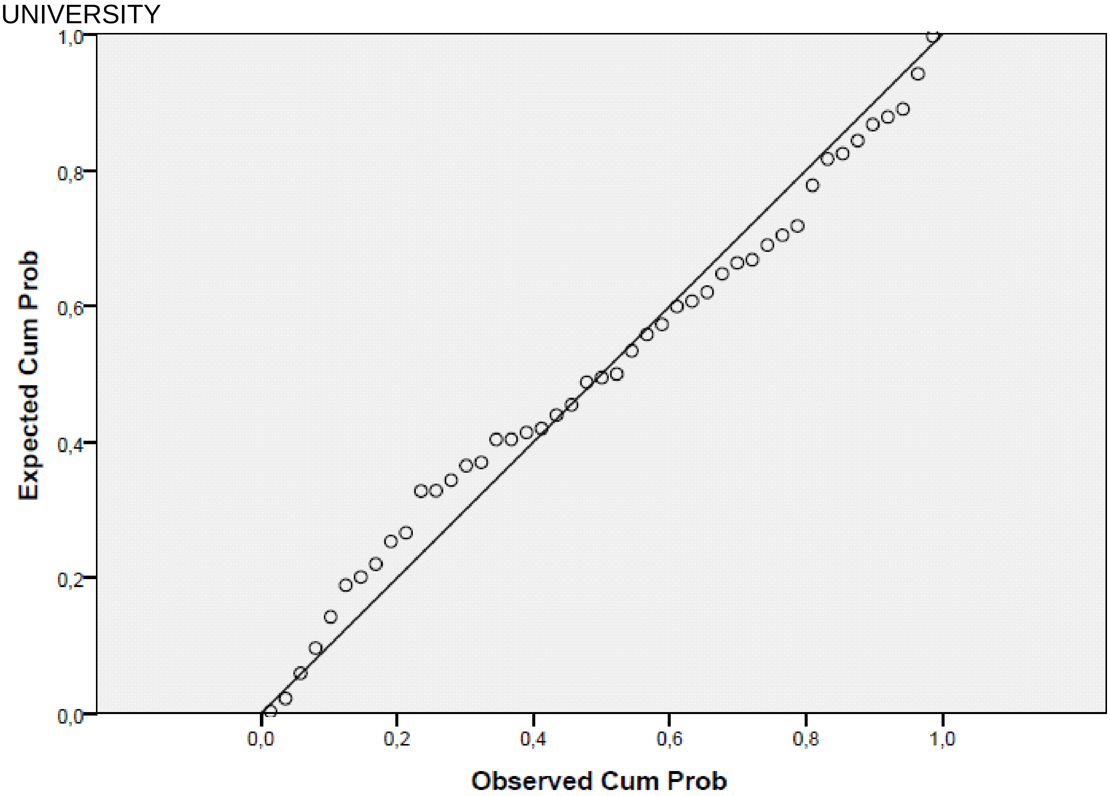


Figure A9.5: Normal probability plot of the standardized residual after adding the variable UNIVERSITY



Appendix 10: SPSS output for the Non Randstad Category 6 stations reference class

Table A10.1: Part of the correlation matrix from SPSS

		CII_RES_RATE	POP_TOTAL	POP_15T65
CROSSINGS	Pearson Correlation	-,058	,904**	,893**
	Sig. (2-tailed)	,614	,000	,000
	N	78	78	78
BUILTUP_P	Pearson Correlation	,084	,880**	,880**
	Sig. (2-tailed)	,466	,000	,000
	N	78	78	78
CII_RES_RATE	Pearson Correlation	1	-,108	-,106
	Sig. (2-tailed)		,348	,357
	N	78	78	78
POP_TOTAL	Pearson Correlation	-,108	1	,999**
	Sig. (2-tailed)	,348		,000
	N	78	78	78
POP_15T65	Pearson Correlation	-,106	,999**	1
	Sig. (2-tailed)	,357	,000	
	N	78	78	78

STOP_60, CROSSINGS, POP_15T65, HH_TOTAL, ONEP_HH, MP_HH_NC, MP_HH_WC, and CAR_OWN are the variables excluded as a consequence of multicollinearity. URBANIZATION_1 and TYPE_SCORE are not included because all cases have the same score for these variables.

Table A10.2: Casewise diagnostics of the first run

Case Number	Std. Residual	N2006	Predicted Value	Residual
7	-2,208	1778	3008,96	-1230,955
8	4,047	5486	3229,67	2256,335
14	2,481	3010	1626,45	1383,548

Table A10.3: Casewise diagnostics of the second run

Case Number	Std. Residual	N2006	Predicted Value	Residual
10	-2,437	370	1492,21	-1122,205
13	2,646	3010	1791,78	1218,221

The first run showed Boxtel as a case with a standardized residual of more than 4. It is excluded for the second run, which yields no outliers. The model is therefore accepted.

Table A10.4: Coefficients

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
5	(Constant)	-989,800	183,431		-5,396	,000
	POP_TOTAL	,125	,013	,545	9,911	,000
	FREQUENCY_oS	250,830	46,584	,387	5,384	,000
	BTM_TOTAL	59,614	14,970	,230	3,982	,000
	STOP_30	-105,205	30,130	-,212	-3,492	,001
	TOT_60	17,042	6,563	,173	2,597	,012

The coefficients from the second run, shown in table A10.4, lead to the following function:

Y = -989.8 + 0.125 * POP_TOTAL + 250.8 * REQUENCY_oS + 59.6 * BTM_TOTAL -105.2 * STOP_30 + 17.0 * TOT_60

The model includes five variables, of which POP_TOTAL is the most influential. All variables are significant at, at least, the 95 percent level (see table A10.4). The model is significant at the 0.000 percent level (see table A10.5). The model has an R^2 of 0.823, which means it explains 82.3 percent of the variance in the dependent variable.

Table A10.5: ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
5	Regression	66949931,391	5	13389986,278	63,162	,000[e]
	Residual	14415713,649	68	211995,789		
	Total	81365645,041	73			

Table A10.6: Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,741[a]	,549	,542	714,169
2	,869[b]	,755	,748	530,309
3	,885[c]	,783	,774	501,665
4	,897[d]	,805	,794	479,210
5	,907[e]	,823	,810	460,430

Figure A10.1: Histogram of the normal distribution of the residuals

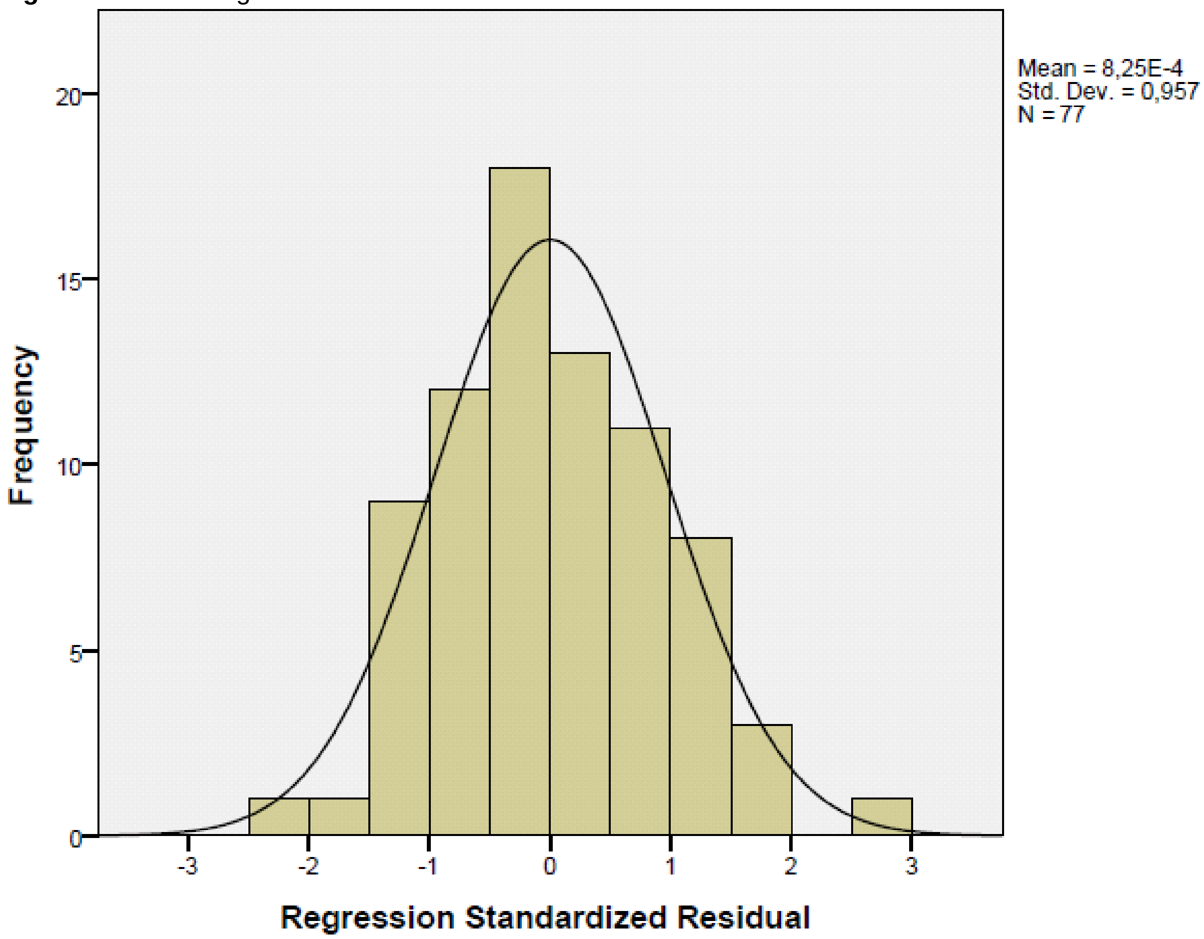


Figure A10.2: Normal probability plot of the standardized residual

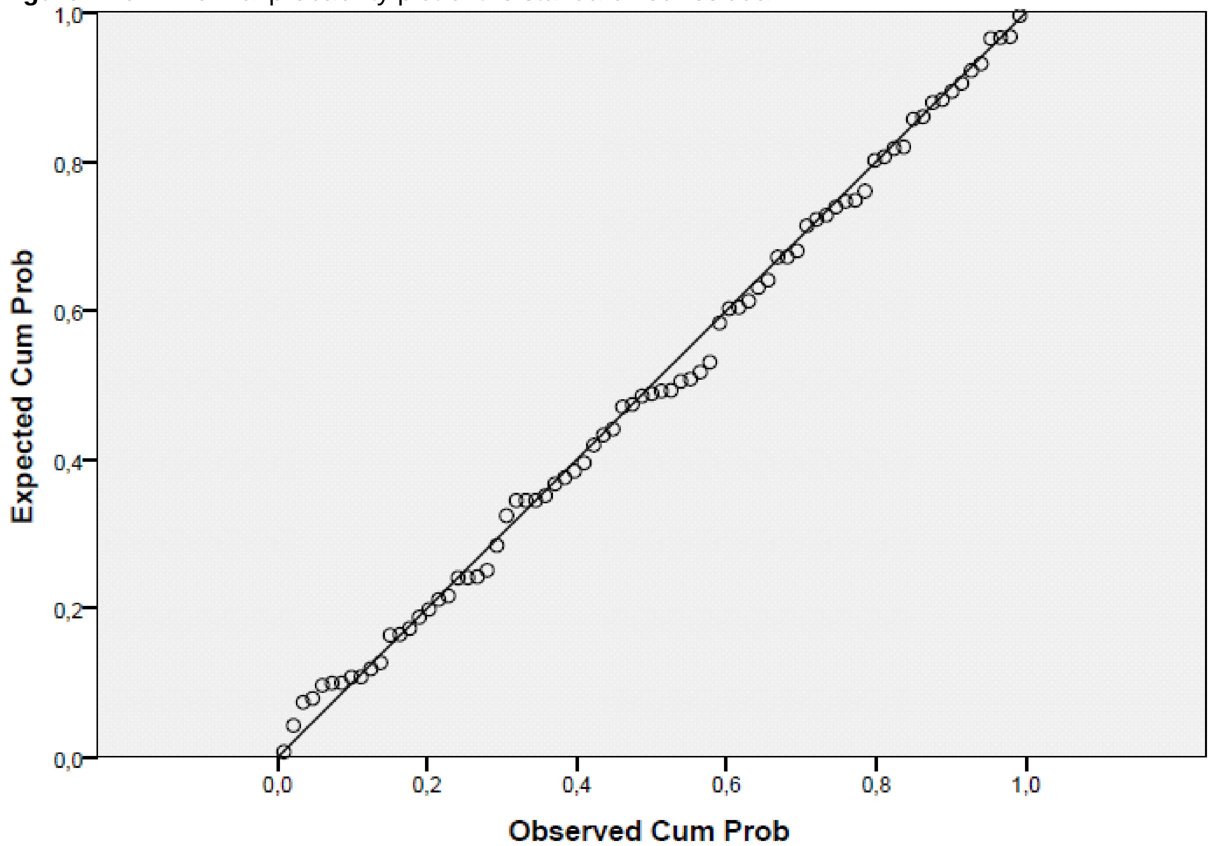
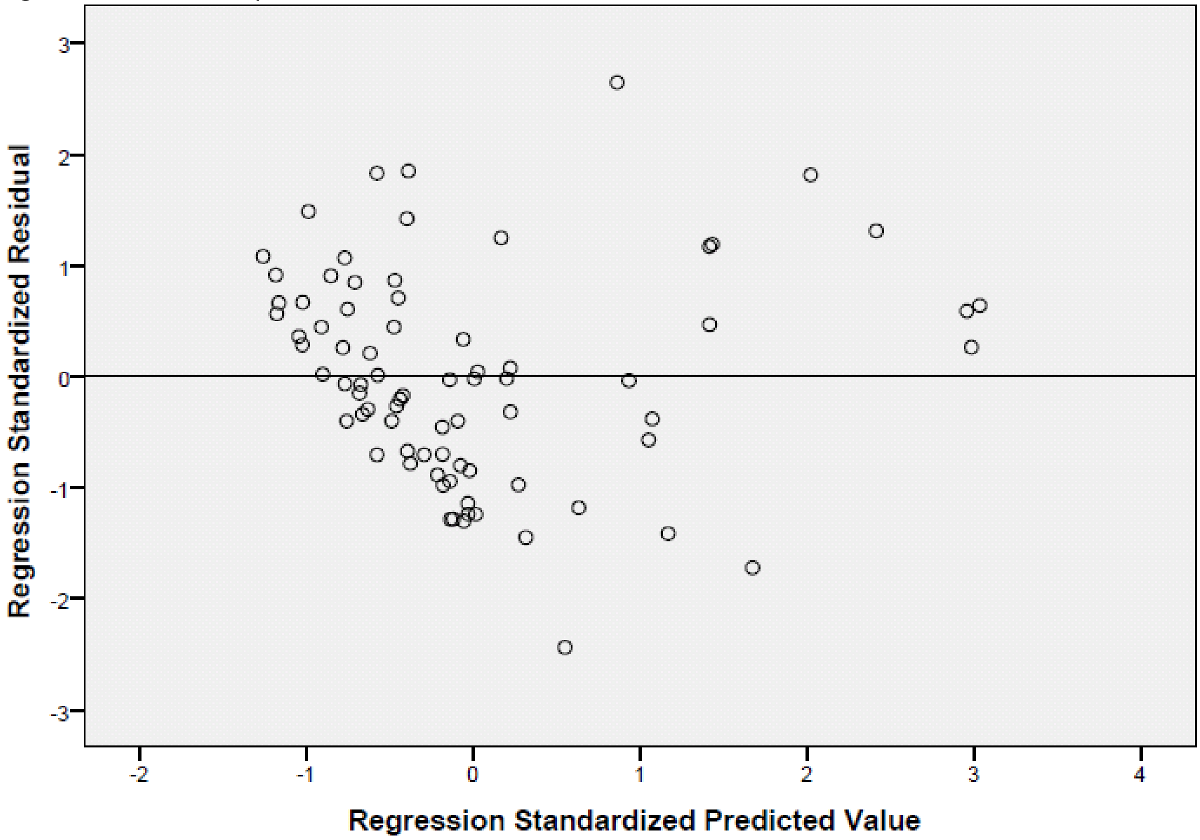


Figure A10.3: Scatterplot of the distribution of the standardized residuals



Appendix 11: SPSS output for the Randstad Category 5 stations reference class

Table A11.1: Part of the correlation matrix from SPSS

		CII_RES_RATE	POP_TOTAL	POP_15T65
HH_TOTAL	Pearson Correlation	-,177	,982**	,987**
	Sig. (2-tailed)	,262	,000	,000
	N	42	42	42
ONEP_HH	Pearson Correlation	-,106	,920**	,936**
	Sig. (2-tailed)	,503	,000	,000
	N	42	42	42
MP_HH_NC	Pearson Correlation	-,274	,963**	,956**
	Sig. (2-tailed)	,079	,000	,000
	N	42	42	42
MP_HH_WC	Pearson Correlation	-,235	,955**	,941**
	Sig. (2-tailed)	,134	,000	,000
	N	42	42	42
CAR_OWN	Pearson Correlation	-,237	,938**	,926**
	Sig. (2-tailed)	,130	,000	,000
	N	42	42	42

TOT_60, STOP_30, POP_15T65, POP_ALLOCH, HH_TOTAL, ONEP_HH, MP_HH_NC, MP_HH_WC, and CAR_OWN are the variables that are excluded here for multicollinearity (see table A11.1). The Randstad variable is not taken into account because all stations are located within it, so this variable does not diversify. This also applies to the variables CATEGORY_5 and CATEGORY_6, since all stations in this reference class are category 5 stations.

Table A11.2: Casewise diagnostics

Case Number	Std. Residual	N2006	Predicted Value	Residual
19	-2,646	4126	12713,15	-8587,148
26	2,487	39555	31482,22	8072,779

As shown by table A11.2, the first run does not bring forward any cases with a standardized residual over 4. The model is thus accepted. The coefficients of the model, which are presented in table A11.3, make up the following function:

Y = -6481.6 + 816.8 * FREQUENCY_oS + 3658.4 * BTM_BINAIR

The model exists of two variables, both significant at, at least, the 95 percent level.

Table A11.3: Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
2	(Constant)	-6481,562	1534,292		-4,224	,000
	FREQUENCY_oS	816,796	78,005	,841	10,471	,000
	BTM_BINAIR	3658,344	1431,198	,205	2,556	,015

Table A11.4: ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,152E9	1	1,152E9	96,107	,000[a]
	Residual	4,797E8	40	11991379,981		
	Total	1,632E9	41			
2	Regression	1,221E9	2	6,106E8	57,968	,000[b]
	Residual	4,108E8	39	10534032,936		
	Total	1,632E9	41			

Table A11.5: Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,840[a]	,706	,699	3462,857
2	,865[b]	,748	,735	3245,617

Judging by the beta coefficient from table A11.3 and the R^2 of the first part of the MR analysis (displayed in table A11.5), FREQUNCY_oS is by far the most important variable when explaining ridership for this reference class. The complete model is significant at the 0.000 percent level (see table A11.4) and explains 74.8 percent of the variance in the dependent variable (see table A11.5).

Figure A11.1: Histogram of the normal distribution of the residuals

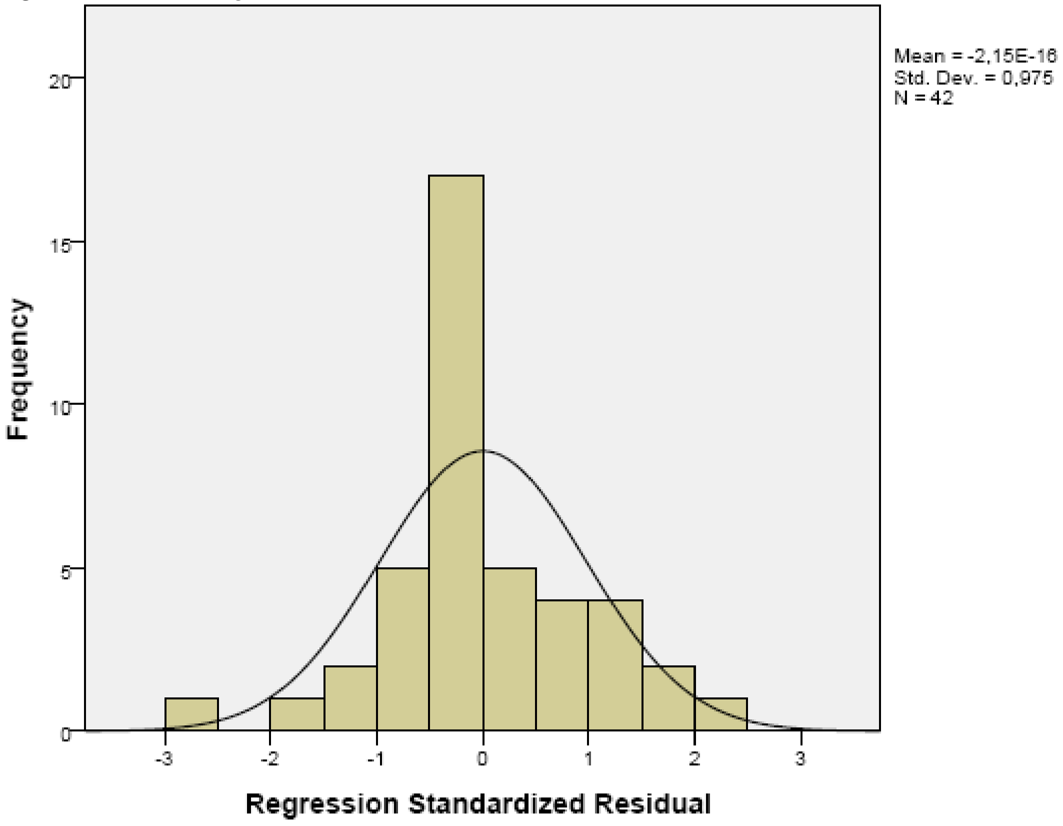


Figure A11.2: Normal probability plot of the standardized residual

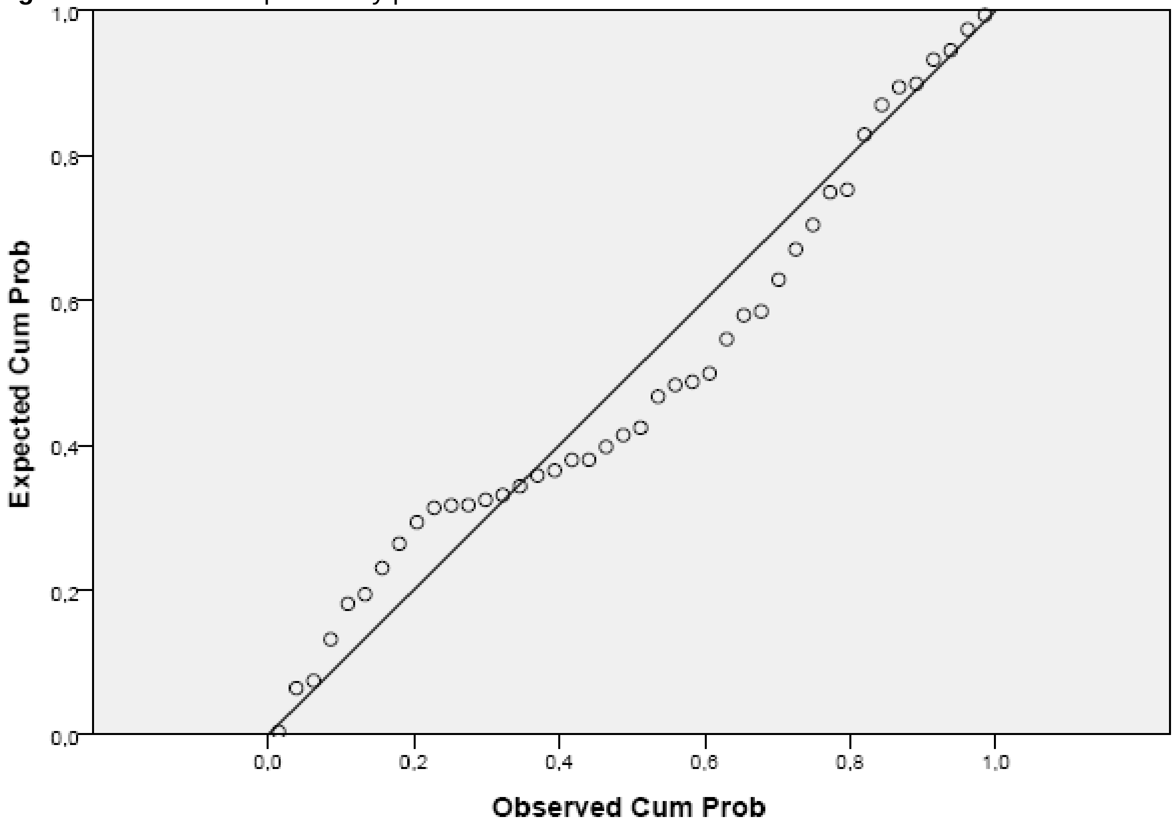
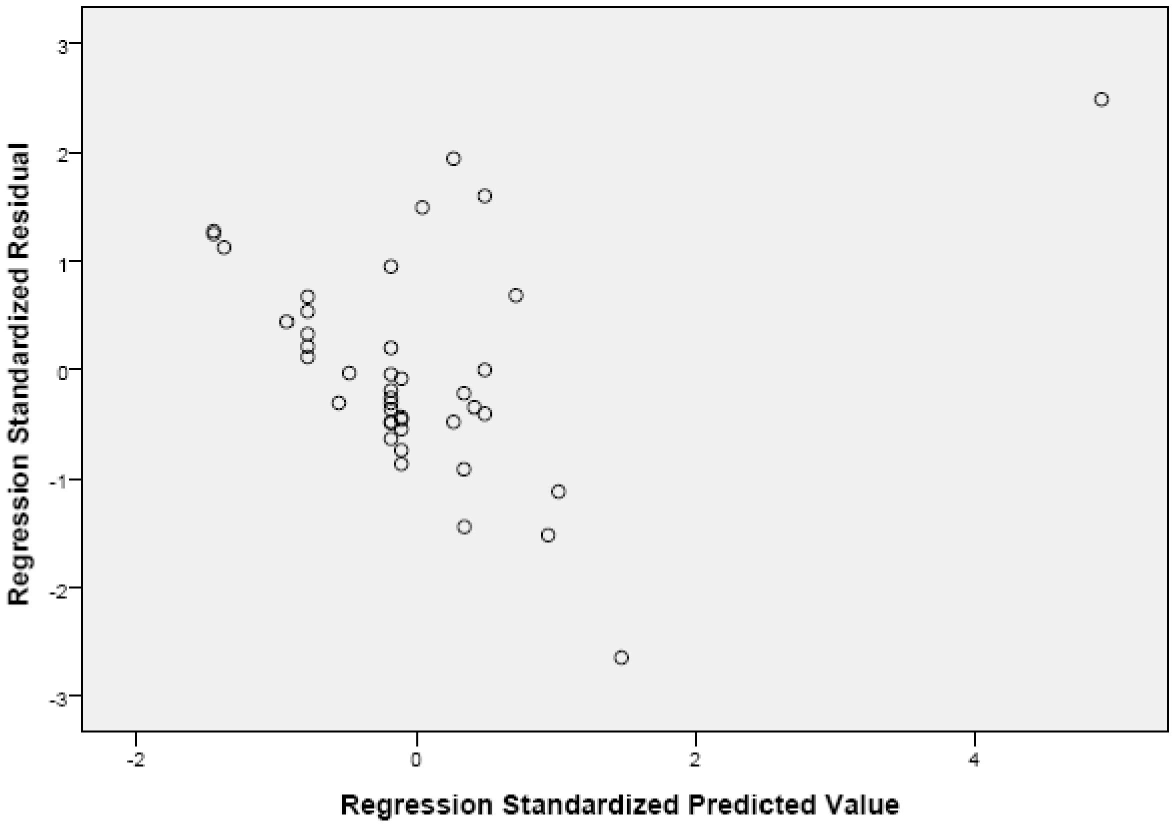


Figure A11.3: Scatterplot of the distribution of the standardized residuals



Because the model only includes two variables and POP_TOTAL is not among those, it is attempted to improve, at least theoretically, the model by adding the latter variable through the Stepwise option. This does, nevertheless, not lead to a significant increase in the model's explanatory and forecasting capabilities (see table A11.6 and A11.7). The variable is far from significant and only leads to an increase in the model's R^2 of 0.1 percent.

Table A11.6: Model summary when POP_TOTAL is included

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,865[a]	,749	,729	3283,998

Table A11.7: Coefficients when POP_TOTAL is included

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-6320,536	1639,119		-3,856	,000
	FREQUENCY_oS	822,674	81,229	,847	10,128	,000
	POP_TOTAL	-,011	,035	-,026	-,306	,761
	BTM_BINAIR	3722,806	1463,350	,209	2,544	,015

Appendix 12: Planned and proposed stations as of 2011

Name	Year	Status	Remarks
's-Hertogenbosch Avenue 2			
's-Hertogenbosch Maaspoort			
Almere Poort	2012		
Arnhem Business Park			
Arnhem Pley			
Assen Zuid			
Avantis			
Barneveld Zuid			
Belfeld			
Bergen op Zoom Zuid			
Berkel-Enschot			For now nonviable
Binckhorst			
BleiZo	2012		
Boskoop		Relocation	
Boven-Hardinxveld	2012		
Breda Oost			
Bunnik		Relocation	
Deventer Noord			
Deventer Zuid-Epse			
Dordrecht Copernicusweg			
Drachten			Proposed track Heerenveen – Groningen
Dronten	2012		New track Lelystad – Zwolle
Echteld			
Eindhoven Noord			Also referred to as Eindhoven Acht or Airport
Eindhoven Strijp S		Vervanging	Instead of Eindhoven Beukenlaan
Emmen Zuid	Opened	Vervanging	Instead of Emmen Brageres
Giessendam De Blauwe Zoom	Opened		
Gorinchem Papland	2012		
Gouda Gouweknoop			
Groningen Europapark		Relocation	
Grubbenvorst			
Haarlem West			
Halfweg-Zwanenburg	2012		
Hengelo Gezondheidspark	2012		
Hoevelaken	2012		
Hoogezand Centrum			
Kampen Zuid	2012		New track Lelystad – Zwolle
Kerkrade De Vink			
Laren-Almen			
Leerdam West	2012		
Leeuwarden Werpsterhoek			
Leiden De Mors			
Leiden Merenwijk			
Lelystad Zuid			
Maartensdijk		Replacement	Instead of Hollandsche Rading
Maastricht Noord	2012		Also referred to as Maastricht Beatrixhaven
Nieuwerkerk ad IJssel			
Nijkerk Corlaer			
Nijmegen Goffert			

Name	Year	Status	Remarks
Oss Oost-Berghem			
Roermond Noord			
Sappemeer			
Sassenheim	Opened		
Schiedam Spaland			Also referred to as Schiedam Kethel
Sliedrecht Baanhoek	Opened		
Sneek Harinxmaland			
Spekholzerheide			
Stadskanaal			
Staphorst			
Utrecht Leidsche Rijn	2012		
Utrecht Majella			
Utrecht Vaartsche Rijn			
Valburg			
Veendam	Opened		
Venlo Zuid			
Vlissingen		Replacement	
Westervoort	Opened		
Zevenaar Oost			
Zwijndrecht Bakestein			
Zwolle Zuid			