

Estimating Railway Ridership

DEMAND FOR NEW RAILWAY STATIONS IN THE
NETHERLANDS

TSJIBBE HARTHOLT
S1496352

COMMITTEE:

K. GEURS (Chairman)	University of Twente
L. LA PAIX PUELLO	University of Twente
T. BRANDS	Goudappel Coffeng

UNIVERSITY OF TWENTE.

adviseurs
mobiliteit
**Goudappel
Coffeng**

I. SUMMARY

Demand estimation for new railway stations is an essential step in determining the feasibility of a new proposed railway stations. Multiple demand estimation models already exist. However these are not always accurate or freely available for use. Therefore a new demand estimation model was developed which is able to provide rail ridership estimations.

Main question of this thesis that will be answered is:

How can the daily number of passengers of a new train station be forecasted on the basis of departure station choice and network accessibility?

Aim is to estimate a demand estimation model which is valid for the whole of the Netherlands and focusses on proposed sprinter train stations.

Factors determining total rail ridership

Rail ridership can be determined by three main factors:

- Built environment factors
- Socio-economic factors
- Network dependent factors

Built environment factors are factors that describe the situation in the direct environment of the station. A subdivision can be made into station environment factors based on the three d's as described by Cervero and Knockel-man (1997):

- Density: Describing the amount of activities in the proximity of the station. This could be the e.g. number of jobs, number of students, shops or total population.
- Diversity: describing the diversity of the activities that take place in the proximity of the station.
- Design: variables describing the properties of a station (area) as a direct consequence of its design. E.g. the accessibility by bike (bicycle parking available), design of the station itself (architecture) or perceived safety.

The socio-economic variables are mainly adding an additional layer to the density variables. They give additional information on for example income, employment, age, or car ownership which can increase or decrease the probability a person will use the train.

Network dependent variables describe the connectivity of the station with the other station in the network. This can be described with variables such as frequency, number of lines, intercity service available or an accessibility index. Secondly, network dependent variables can also describe the quality of the potential feeder modes such as the frequency and number of lines for bus, tram and metro or the availability of a park & ride. In total 147 variables have been categorized and tested for their explanatory value.

Effects of a new station

The opening of a new train station can have several effects. Generally it is assumed a new station will mainly attract new passengers. Because of increased rail accessibility (closer station proximity) after the opening of a new station, this will be most likely the case for some people. However, this increased rail accessibility will also cause an abstraction of demand from existing stations. A part of the passengers using the new station are therefore existing train users. Only their station preference has changed.

Finally, a new station can also cause a decrease of passengers elsewhere along the line because of the (slightly) longer travel time. An additional stop a train has to make will increase overall travel time by three minutes on average. Existing passengers might therefore decide to use another mode of transport due to this increase in travel time.

Methods available to estimate travel demand

Two main types of demand estimation have been identified:

- Aggregated demand estimation
- Disaggregated demand estimation

Aggregated demand estimation is usually based on regression analysis according to the formula:

$$Y_i = \beta_0 + \sum_k \beta_k \beta_{ik} + \varepsilon_i$$

With parameters:	Y_i	the total number of predicted passengers
	B_0	The constant or intercept
	B_k	Estimated parameter for variable k
	B_{ik}	variable value i for variable k
	ε_i	error term for variable i

This model is commonly used since no disaggregated trip data is needed and is relatively easy to apply. However, regression models are sensitive for the quality of the variables used and potential outliers in the dataset. In order to further improve a regression model several additional actions can be performed:

- Reference class forecasting: With reference class forecasting all cases are assigned to separate classes together with other similar cases. This will allow for the estimation of separate models adjusted to the reference classes.
- The use of network distances: By using the network distances instead of Euclidian distances, the accuracy of variables such as the total population the proximity of a station will be improved. The problem of barriers in the landscape such as rivers, highways and the railway line itself limiting the actual catchment area will be solved using this method. ((Upchurch , Kuby, Zoldak, & Barranda, 2004), (O'Neill, Douglas , & JaChing, 1992), (Horner & Murray, 2004).)
- Distance decay modelling: In several cases it has been observed that people living further away from the station have a lower probability of using the train (Keijer & Rietveld, 2000). Adjusting to this affect with the use of distance decay can therefore improve several variables such as total population) significantly (Gutiérrez et al, 2011).
- The use of geo-weighted regression allows for a geographic variation in the constants of regression model. Therefore a geo-weighted model can adjust for region differences in the sensitivity of certain variables ((Blainey & Mulley, 2013).

Disaggregated demand estimation is usually based on disaggregated trip data. The need for this kind of data makes it harder to apply this type of model. However this type of model is better suited to estimate effects on station choice and competition between stations. It is often applied with the use of a multinomial (or nested) logit model. Such a model will offer multiple alternatives (stations). Based on the unique situation of each case a utility will be assigned to each of the choices. The probability of choosing a choice is then calculated based on these utilities.

Research method

In this research a combination of these two methods will be used: A multinomial station choice model will be used to improve variables before they are used in a regression analysis.

Furthermore an accessibility indicator and distance decay function are estimated to be used as model input as well.

Accessibility Indicator

The position of the station in relation to the rest of the network has proven to be an important factor in rail demand estimation. In this research an accessibility indicator was estimated to include this aspect in this model as well. These indicators were based on a trip distribution model estimated in Omnitrans. In total three indicators were estimated. The final index score is normalized from 0 till 1.

For example the closeness centrality index (CCI) was estimated as:

$$CCI_i = \sum_{ij} (\delta_{cij} * D_j * \frac{1}{C_{ij} + 1})$$

With parameters:

CCI_i	The closeness Centrality Index of station i
δ_{cij}	The probability of taking a trip from station i to j
D_j	The total number of passengers arriving at station j
C_{ij}	The number of transfers needed to get from i to j

Distance decay functions

Based on survey data conducted in the province of South-Holland distance decay functions were estimated. The functions are separately estimated per station type on the access side and separately for sprinter and intercity stations on the egress side. Multiple function types have been tested but a logarithmic function type proved to have the best fit.

The largest difference can be observed between intercity (type 1 & 2) and sprinter stations (type 3 till 6) with intercity stations having a considerable larger catchment area and trip attractively. However, type 1 intercity stations seem to have a slightly larger catchment area than a type 2 station. At the same time type 5 sprinter stations have the smallest catchment areas.

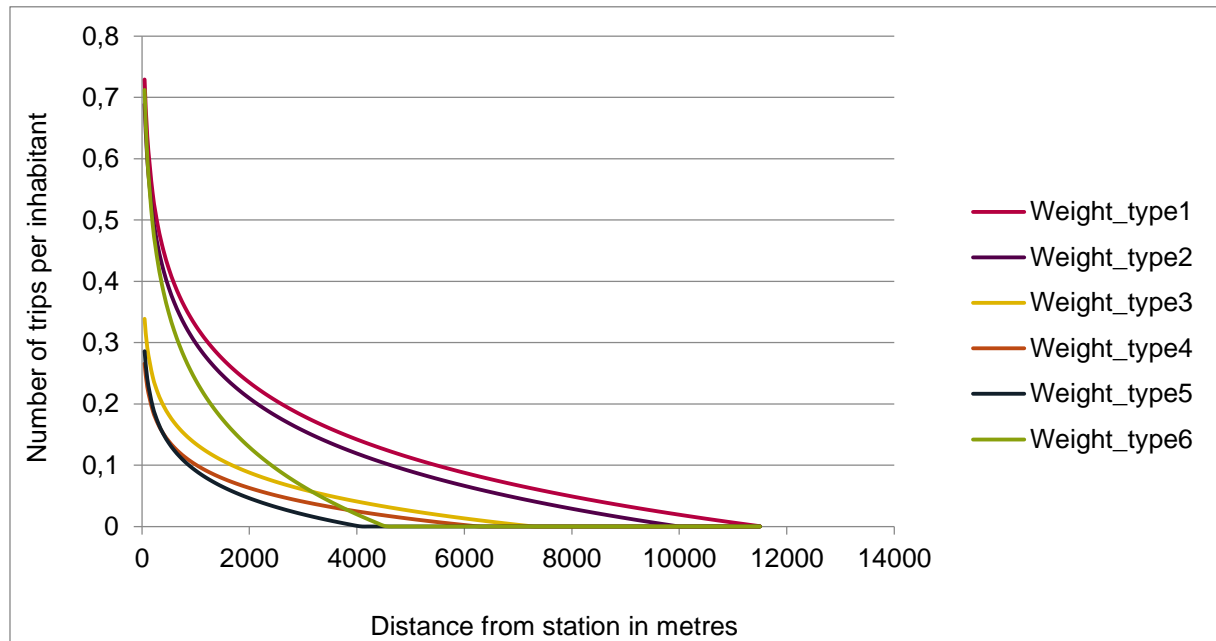


Figure 1: Distance decay functions per station type on the access side of the trip

Station choice model

Also a multinomial station choice model was estimated based on survey data and the use of Biogeme. The final station choice model was based on a choice set consisting of two closest intercity stations and two closest sprinter stations. Variables included in the model were frequency, availability of guarded bicycle parking, number of BTM lines connecting the station and distance.

Regression analysis

A regression analysis was performed on the basis of variables adjusted with the distance decay functions and the station choice model resulting in the total potential of train trips from the number of jobs, student places and total population. Furthermore the closeness centrality indexes along with several other variables were included as well. Six different models have been estimated. Two of these models are valid for all sprinter stations, four models are type specific models based on the reference classes: regional and main line models (Table 1).

Table 1: Overview of all estimated regression models

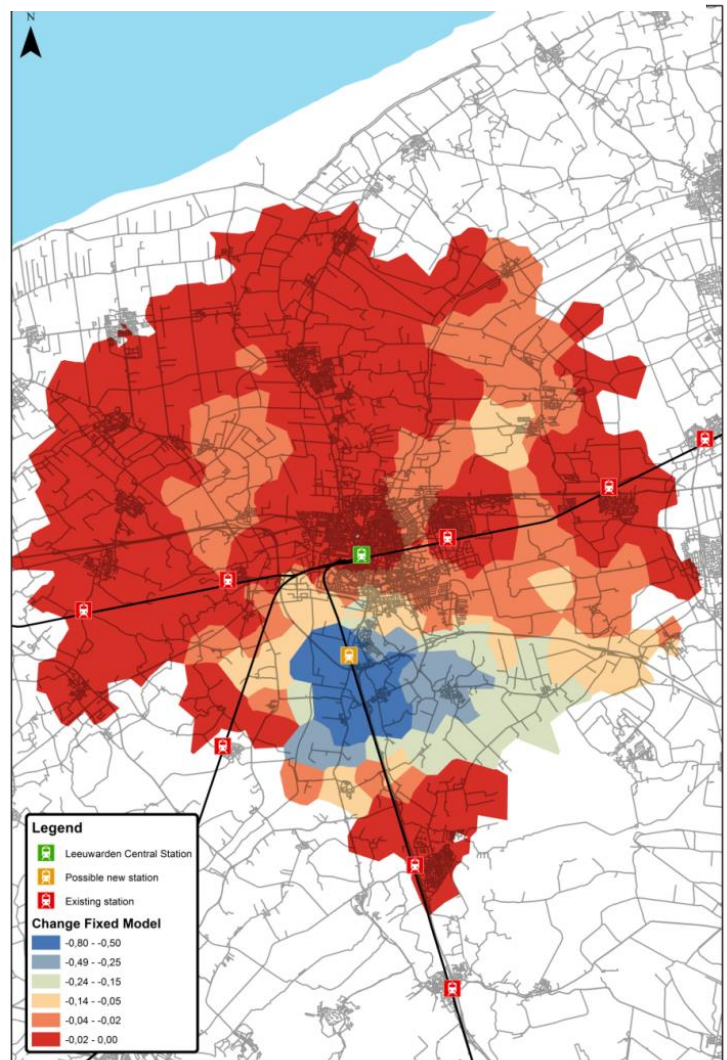
	General Basic	General extensive	Regional basic	Regional extensive	Main line basic	Main line extensive
Cases	307	307	119	119	191	191
R²	0,837	0,871	0,728	0,789	0,798	0,819
Std. Error of the Estimate	1005	894	556	489	1193	1140

Application & discussion of the model

Application of the model can give a demand estimation of the new station. The effects of demand abstraction of the new station on existing stations can be estimated with the station choice model (see figure 2). When applied the two general model will give the most accurate results. The type specific models will give the least accurate results.

Limit of this model is the fact it does not incorporate mode choice as part of the demand estimation. Furthermore, only station type based decay functions have been tested. However, decay function based on access mode choice could be very useful as well, especially in combination with the attractiveness of each station for each mode.

Figure 2: demand abstraction of Leeuwarden as a result of the opening of Leeuwarden-Werpsterhoek



II. TABLE OF CONTENTS

I. Summary	2
1. Introduction	10
1.1 Problem definition.....	10
1.2 Objectives.....	12
1.3 Research questions.....	12
2. Theoretical Framework.....	13
2.1 Factors determining Basic Rail demand	13
2.2 Effects of opening a new station	21
2.3 Modelling New Stations.....	22
2.4 Stations in the Dutch practice	30
3. Methodology & Data	32
3.1 Research Approach.....	32
3.2 Analytical framework	33
3.3 Modelling steps	35
3.4 Data	36
3.5 Model Validation.....	41
4. Model Estimation	43
4.1 Accesability indicator.....	43
4.2 Distance Decay Functions	49
4.3 Station Choice Model	57
4.4 Initial Station Potential.....	65
4.5 Corrolations	67
4.6 Regression Models.....	72
4.7 Geoweighted Calibration.....	79
4.8 Model validation	82
4.9 Reliability of results	86
4.10 Model Application	90
5. Discussion	93
5.1 The use of the rail accesability indicator	93
5.2 Station Potential & station choice model.....	93
5.3 Regression Models.....	94
6. Conclusions	96
6.1 Research questions.....	96
References	99
Appendices	103
Appendix 1: proposed stations in the Netherlands	103
Appendix 2: Complete list of all variables	105

Appendix 3: Overview of MNL station choice model 1.....	107
Appendix 4: Potential for sprinter stations.....	108
Appendix 5: Inter-Correlation between variables	112
Appendix 6: Correlation of Final regression models (minus Outliers).....	113
Appendix 7: Overview of all stations with actual and estimated demand.	114
Appendix 8: Overview of all validation stations and their estimated ridership for all models.....	119
Appendix 9: Selection of proposed stations with ridership estimation and error margins	120
List of figures & tables	122
List of Tables	122
List of figures	123

1. INTRODUCTION

1.1 PROBLEM DEFINITION

The Dutch railway network is one of the densest and heavily used networks in the world right after Japan and Switzerland. The total amount of passenger kilometres increased from 14 billion in 2004 to 17 billion in 2013. Moreover, several new stations are opened almost every year. In the last 20 years 40 new stations have been opened in total.

The initiative for a new station can come from local governments such as provinces and municipalities or city regions. The rail operator (e.g. NS, Arriva, and Syntus) will then make an estimation of the feasibility of a new station based on the expected amount of passengers. However, there is often a difference in perspective on the feasibility of a new station. Rail operators can be cautious for opening new stations as the expected number of additional passengers is not always sufficient. It is common that the local governments are expecting larger benefits from opening a new station than the railway operator. Therefore the process of opening a new station is often a long and difficult process and might take several years to even decades depending on the expected feasibility of the station.

Secondly, in order to be eligible for funding by the national government for setting up a new station, the proposal has to meet certain requirements. Firstly there needs to be a guarantee that the transport operator will serve the new station in the timetable. Secondly the station should have a fitting business case concerning the station itself as well as the station environment. The financial costs should be completely covered. If these requirements are met the new station can receive a subsidy of a maximum of 6.5 million euros (Ministry of I&M, 2014)

Demand forecasting errors

Worldwide, almost 9 out of 10 rail projects including new infrastructure, stations, and high speed railway lines, have an overestimated demand upon completion. On average this overestimation is about 106% of the actual flow of users. For 50% of the road projects this overestimation is only about 20% of the actual use (Flyvbjerg, et al., 2005). It also appeared that out of 58 rail projects in the dataset used, the average costs escalation was 44.7%. Compared to other project types this cost escalation was much lower such as fixed links with 33.8% escalation and roads with 20.4% (Bent Flyvbjerg, et al., 2003).

Although academic research on the comparison between actual and predicted demand in a Dutch context is missing, it appears from data of the 2009 document ‘*toepassing norm nieuwe in- en uitstappers bij nieuwe stations*’ that demand prediction (using the demand estimation PINO from Dutch Railways) in the Netherlands is, likewise as in the research of Flyvbjerg, not always close to actual demand. In table 2 a comparison is made between the predicted and actual travel demand. This comparison is based on stations opened in the Netherlands between 2003 and 2007. All stations are compared with the actual travel demand in the year 2009 and 2013, the most recent year of which travel demand data is available. The average overestimation based on data from this document is about 31% in 2009. Stations which have been replaced, that were only temporary or those still under construction are not taken into account.

It can be seen in table 2 that the current predictions tend to overestimate the ridership on the short term. However in the middle long term demand can still grow, causing the average estimation error to decline to only -6.3%. However on an individual station level difference between prediction and actual demand can still be rather large as the average size of the error (positive or negative) only declines from 34.4 to 23.0 percent. It must be noted that on the longer term, predictions become less valuable as other factors which can change in time are not taken into account in the demand model. And as rail demand on a national scale has been growing in the period 2009-2013 it makes sense that this trend is also to be seen in the daily boarding at the train stations in this list.

Table 2: Comparison between predicted and actual ridership demand

Station	Year of opening	Predicted (PINO)	Actual (2009)	Actual (2013)	% Error (2009)	% Error (2013)
Amersfoort Vathorst	2006	2500	1840	2559	-26.4	2,4
Tiel Passewaaij	2007	1100	1230	1269	11.8	15,4
Utrecht Zuilen	2007	2000	1397	1918	-30.2	-4,1
Amsterdam Holendrecht	2008	3250	3111	3176	-4.3	-2,3
Apeldoorn de Maten	2006	1750	636	1040	-63.7	-40,6
Apeldoorn Ossenveld	2006	1500	773	n.a.	-48.5	-
Gaanderen	2006	550-750	339	n.a.	-47.8	-
Voorst-Empe	2006	350	288	n.a.	-17.7	-
Twello	2006	1750	1330	1554	-24.0	-11,2
Purmerend Weidevenne	2007	2000-2250	1578	1646	-25.7	-22,5
Heerlen de Kissel	2007	800-1200	419	n.a.	-58.1	-
Eygelshoven Markt	2007	400	149	n.a.	-62.8	-
Tilburg Reeshof	2003	1600	1838	2563	14.9	60,2
Almere Oostervaarders	2004	3500	3439	4285	-1.7	22,4
Den Haag Ypenburg	2005	2150	1327	1801	-38.3	-16,2
Arnhem Zuid	2005	3900	1945	2790	-50.1	-28,5
Helmond Brandevoort	2006	2050	833	1021	-59.4	-50,2
Average Error					-31.3	-6.3

The causes for these overestimations in rail projects are ascribed to two main reasons: “uncertainty about trip distribution” and “deliberately slanted forecasts” (Flyvbjerg, et al., 2005). The first reason might be because older datasets are used to calibrate the model. Levels of “rail patronage might therefore be over (or under-) estimated” according to Flyvbjerg et al. (2005).

The second reason however is an error which might be subconsciously (optimism bias) or even deliberately put into the forecast. By overestimating the forecasts it is more likely that the project will be build. This overestimation of demand in combination with an underestimation of the societal costs can cause serious welfare reductions as money which could be spend more useful and effective elsewhere is invested in the wrong projects on the basis of false forecasts.

Conclusion is that rail demand estimations at individual stations could be more accurate. Over- or underestimations of more than 20% are no exceptions. Therefore there is room to improve these demand estimations and improve decision-making as with the current method stations are being built which would not have been built if a better forecast would have been made.

Unaccounted ridership effects

Besides errors in the total demand estimation, local ridership effects can have a large impact as well, even when we would be able to perfectly predict the ridership of a new station. Since the goal of opening a new stations often to increase the share of people traveling by train, in reality passengers using a new station might be abstracted from other stations. Opening a new station might only decrease the efficiency of the network in that case.

Secondly, current demand models do not always take into account the fact that new stations are often local stations which offer a lower service levels than intercity stations. Therefore passengers might prefer the intercity stations instead of the (new) local station. These competition effects between stations can have a large impact for the actual ridership as well. The model of the Dutch railways (PINO) is not taking these competition effects into account in a realistic way. Based on PINO, the catchment area of the stations is divided on an all-or-nothing based approach between the two overlapping stations based on frequency. In reality however it can be assumed that there is not a clear border between the catchment areas of two stations.

1.2 OBJECTIVES

Goal of this thesis is:

To develop a demand forecasting method which is able to provide ridership estimations of new sprinter train stations based on station choice, network accessibility and network effects.

After application of this new method it will give an overview of the basic feasibility of a new station. As this method also takes into account the effects on other stations, it will give a better overview compared to methods only reviewing the total number of expected passengers. Also the number of newly attracted rail passengers should be estimated, making this method is more useful in order to test if certain policy goals will actually be achieved by taking the measure of opening a new sprinter station.

1.3 RESEARCH QUESTIONS

In order to reach the goal of this thesis the following main question will be answered:

How can the daily number of passengers of a new train station be forecasted based on station choice and network accessibility?

Before a station can be evaluated there is a need for a clear understanding of what is generating rail demand, by what factors it is affected and how it can be modelled. Therefore the following sub-questions to be answered before making the model have been formulated:

1. *Which factors determine total ridership of a train station?*
2. *What is the effect of a new train station on departure station choice?*
3. *Which methods are available to estimate travel demand?*

When method and model types are known, there are some practical implications which could affect the final model quality:

4. *How do station specific variables (such as station type, -quality, and – facilities) in the Netherlands impact the station catchment area?*
5. *How will network specific variables (such as reliability, accessibility and service level) influencing passenger demand at train stations?*
6. *How is competition between stations included and how is this influencing the total ridership demand*

When final model has been generated the following question should be answered:

7. *What is the explanatory power of the model in predicting future travel demand?*

From the completed rail demand model it could then be expected that it can estimate demand for new sprinter train stations in the Netherlands in an accurate way with known error margins.

2. THEORETICAL FRAMEWORK

Incentives for new train stations

There can be multiple reasons why new train stations are being opened. In practice it is often not only one reason but there are multiple incentives for opening a new train station. However the main goal is in most cases to attract more rail passengers as this is considered a more sustainable way of transport than car. On longer distances train travel should even compete with air travel with the use of high speed rail lines. According to the European white paper on transport (2011) 50% of all intercity passenger and freight journeys should shift from road to rail and water in 2050. In short: there is a big role for rail travel in making the transportation sector more sustainable. A common thought is that new stations can help to achieve this more sustainable transport sector.

Although larger towns and cities generally already have a railway station, there are multiple smaller towns and villages which currently don't have a station. By opening new stations in these towns the goal is usually increase the general accessibility of this area. The town of Dronten for example did not yet have a station until recently. Now the new station Dronten might become a more favourable place to live as commuting to larger cities in the area such as Zwolle has become much easier. The amount of people in Dronten who thinks that this new station offers a better opportunity for a job grew considerably (monitor Hanzelijn, 2014).

However, having a train station in your town also gains a bit of prestige for the local town. Municipalities are therefore not always paying attention on whether or not the station is feasible but tend to have an optimism bias towards the new station by overestimating the positive effects and underestimating the negative effects (Bent Flyvbjerg, et al., 2003).

A final reason which is also closely linked with making the transport sector more sustainable is to reduce congestion and the corresponding externalities on the road network (Adler & van Ommeren, 2015). Especially in the urbanized western area of the Netherlands this is often an important incentive. Stations such as Leidsche-Rijn near Utrecht were developed near large scale developments of new dwellings in order to reduce the car usage in these new neighbourhoods.

Where the reasons for opening a new train stations might be diverse, the effects such a new station can have on local rail demand and station choice are diverse as well. Aim of this chapter is therefore to describe all factors of importance that can influence the demand for rail transport at a new station. To do so, this chapter is divided into five subparts.

The first part will cover the factors influencing basic rail demand. In other words: What variables are generating demand for rail travel? The second part is covering the effects a new station can have in terms of demand for rail transport and how this demand can shift between stations. The third part covers the various modelling techniques to model the demand of new stations based on variables and effects as described in the first two sections. The final part will give an overview of the current state of affairs regarding train stations in the Netherlands including all current proposals of new stations.

2.1 FACTORS DETERMINING BASIC RAIL DEMAND

The very first question that is important to ask when estimating demand for new stations is what factors are influencing demand for rail transport in general. The amount of research done on factors determining ridership is extensive. This means in literature many different types of variables are to be found which hypothetically could affect ridership levels in the Netherlands. In this research ridership factors are decided into three main categories:

1. Built environment factors
2. Socio-economic factors
3. Network dependent factors

Built Environment factors

Main explanatory factor of the ridership is the direct station environment. This can also be summarized by the three D's: density, diversity & design. The more activities (recreational, work, and residential) are taking place in the vicinity of the station the higher the fraction of the people attending these activities will travel by train.

It is only in this category of variables where a division between trips generated by home- and activity-end can be clearly distinguished. A high number of people living near the station will cause for a high number of trips on the home-end. Large healthcare or educational facilities, offices, services and recreation can cause a large number of trips on the destination-end.

This might be important as there are indications that stations mainly receiving journeys on the activity-end of the trip are having a smaller catchment area compared to station at the home-end of the trip (Keijer & Rietveld, 2000). As people near the activity-end of the trip don't always have access to a bicycle or car as they would have on the home-end of the trip. Walking is therefore often the dominant egress mode at the activity-end.

Density

Density is one of three d's commonly ascribed as one of the most important variables for transit oriented development. As already mentioned earlier the more people are living or working in the station area, the greater the share will be of people traveling by train (Keijer & Rietveld, 2000). The fact that density is so important for creating demand is also unveiled in the research of Cervero and Knockelman (1997).

The variable can be measured in multiple ways. Sometimes the total land use for several categories is used (i.e. total commercial land use, total residential land use). In one article a differentiation was made between density of service and commercial land use for example (Sung & Oh, 2010). Better might be to take the developed floor area per land-use function as done in the study of (Sung & Oh, 2010). This way high rise developments, which use relatively few square meters on the ground floor are taken into account in a better way as all square metres of all storeys of the building are counted. Sometimes however a more specific indicator is used such as the amount of jobs or total population in an area. Depending on which density is measured density can help explaining as well as trips on the home side (dwellings, inhabitants) as on the activity side (jobs).

Large institutions which can draw a considerable crowd also should be included in this analysis mostly because of the trips at the activity end of the trip. These institutions can consist of large educational institutes such as large schools and universities. Secondly large leisure activities such as museums, theme parks, malls and other leisure/recreational destinations should be included. The potential effects these institutions can have on ridership are often not covered by only taking the jobs into account these institutions offer. Better is to also incorporate the visitors these facilities attract into the equation if this data is available.

Finally, there are also several types of services which, in large densities, can generate a lot of additional trips. These types of services can consist of shops, restaurants, cafés, bars and hotels and other. They can also be subdivided in for example basic needs shops and occasional needs shops (Carpio-Pinedo, 2014).

Diversity

Diversity is said to be less important for creating demand than density. However a large diversity does allow for a more evenly spread demand over time. A high diversity does for example not only attract commuters going to work but also leisure related journeys. “*Land-use mix (diversity) produces a more balanced demand for public transport over time (reducing differences between peak and off-peak periods) and in space (in terms of direction of flow)*” (Cervero, 2004).

Diversity is measured by taking the surface area of each type of land use and calculates the land-use mix (LUM) with the corresponding formula:

$$\text{Land use mix} = \sum_j \frac{(P_j * \ln(P_j))}{\ln(J)}$$

With parameters:	P	total proportion of land use type j
	j	land use category j
	J	total number of land use categories

An outcome close to 1 means a high diversity, an outcome close to 0 means a low land use diversity. This method was used before in studies of Cervero and Knockelman (1997) among others. It is expected that a high diversity will result in a more even distribution of trips generated by the origin side and trips generated by the activity side.

Design

In the variable category “design” we can allocate variables that describe how well the station is accessible by various modes and how passengers are experiencing traveling by these modes to the station. Traveling by train will become more favourable as the station itself is better accessible by bike and foot. Street density is a good indicator for the accessibility by foot of a location (Zhu & Lee, 2008). In a Dutch context where cycling is an important feeder mode, the density of cycling lanes could be used as an indicator as well. From further literature it also revealed that the density of four way intersections appeared to be a good indicator as well (Sung & Oh, 2010).

The quality of access of the station by foot or by bike is affected both the home-end as well as the activity end of trips. Although it can be suggested that variables determining the quality of the accessibility by bike do have a stronger impact on the home-end of the trip as based on Keijzer and Rietveld (2000) it was mentioned that the bike by far the most dominant access mode on the home-end of the trip.

Other design related factors can be related to the station itself. The way the station is experienced and how it is designed can contribute considerably to the daily amount of passengers using the station. The type and amount of services provided, safety, cleanliness and the (architectural) designs itself are all factors that contribute to the overall station satisfaction.

Cascetta and Cartení (2014) determined many different attributes which are all part of station quality. These attributes can be cleanliness, information availability, security, climate control, architectural/aesthetic quality and several others. Many of these attributes can also be subdivided into a subjective and an objective version of the variable. As for example security can be objectively very high (i.e. because of a low number of crimes) but passengers still might feel very unsafe.

Recent research proved that the overall station quality can have a large impact on the number of travellers. By comparing two metro lines through homogeneous urban areas in Naples it appeared that the architecturally upgraded metro line had a larger catchment area. For the access mode “walking” this meant a catchment area increase of about 400 metres based on access distances retrieved from

questionnaires. The willingness to pay for this line was 35 cents higher for students, and 50 cents higher for commuters (Cascetta & Cartení, 2014).

The quality of certain facilities for passengers can also be a factor in station choice and access mode choice. It was unveiled that improvements in guarded and unguarded bicycle parking at stations in the Netherlands could enlarge the share of cyclists as an access mode to the station. However, the availability of parking in rush hour is one of the most important factors (La Paix Puello & Geurs, 2015).

The profile of a station (does the station attracts mainly trips on the activity or home-end) can also determine the effectiveness of certain station facilities. Trips on the activity-end usually have a higher degree on walking and BTR as access/egress modes contrary to trips on the home-end where bicycle is more often used (Keijer & Rietveld, 2000). This indicates that certain variables do not have the same impact at every type of station.

In short it can be concluded that the way the station looks like and how the station is experienced can make a large difference in the size of the catchment area and ultimately in the total ridership such a station can generate. However these variables are hard to measure objectively and this can only be done by conducting a survey at the stations.

Secondly the facilities such as bike parking, car parking, restaurants, and free internet can contribute to the overall experience. Hereby it does not only count if they are present but also what the quality and availability (during rush hour) of these facilities is. Again a survey amongst users or at least an observation of these facilities would be necessary in order to measure the quality of these facilities.

Socio-economic factors

Socio-economic circumstances can have a great impact on ridership levels. These indicators do not give the amount or density of people in a certain area. Instead they give an additional layer of information about the density in an area. These variables are therefore not main indicators of ridership but can explain the difference between two (in terms of density) similar stations.

The characteristics of a train user

The relation between socio-economic variables and rail ridership can best be explained by dividing train users in two groups:

- Train users by choice
- Captives

(Brown, 1983) (Polzin, et al., 2000)

This categorization of train users is already used for at least 30 years and still is in use in current literature although with the rise of modern technology (such as car sharing apps) the division between captives, users and non-users becomes more a grey area. The division is based on people who are able to travel by another mode if they wanted to but still decide to use the train on one hand. People who have no choice and are therefore forced to use public transport on the other hand (i.e. because they don't have a car or driving license). The reason for being a public transport captive is also often related to a low income, health issues and age (Krizek & El-Geneidy, 2007).

Based on the outcome of the Dutch Railways (NS) customer satisfaction survey carried out between a Monday and Friday in September 2005 it can be estimated that for the Dutch case almost half of the train passenger market consist of non-captive passengers (Givoni & Rietveld, 2007). Captive passengers tend to be less content with the overall travel experience compared with the non-captive group which can be explained by the fact that the captive group also contains people who would rather choose for a car if given the choice (Brons & Rietveld, 2009). The captives are, as they don't have access to a car, relying on public transport, bicycle or walking as access mode to the station. Non-

captives have the opportunity to go by car to the station as an access mode if they choose to travel the main leg of the journey by train.

The distance train users are willing to travel in order to reach the station depends on their access mode and the service offered at this station. It is known that people who live nearby a train station are more inclined to take the train than people who live further away (Keijer & Rietveld, 2000). However, also personal circumstances of the passengers can affect the distance a person is able or willing to travel to a train station.

Another research showed that young people and adults without children, men, immigrants, and public transit captives are willing to walk longer distances and are less sensitive to the effect of distance (García-Palomares, et al., 2013). In research of it appeared also that elderly tend to travel smaller distances (average of 13 kilometres) by train compared to middle aged and young people (average of 16 kilometres) (Akiyama & Okushima, 2009). This group of elderly also tends to avoid transfers more compared to other age groups. However it should be noted this research was done at a metropolitan railway system in Japan and therefore transferability of the results to a Dutch context should be handled with care.

Car ownership is one of the most profound social-economic variables. As stated earlier, there are two types of rail passengers: captives and non-captives. If more households own a car then more people are having a choice between car and train. One would therefore expect that car ownership is a negative factor for rail demand. This relation was also confirmed in literature (Wardman, et al., 2007).

Income is also a variable which can affect ridership. From previous studies it is known that higher income groups generally make less use of public transportation. Therefore the amount of people with a high income can have a negative influence on ridership (Babalik-Sutcliffe, 2002). The amount of students in the catchment area of a station is usually seen as positive for public transportation demand. A positive correlation was found between the percentage of students living nearby and rail demand in the study of Wardman et al. (2007).

The number of renters (contrary to home owners) was used in a study of Kuby et al. (2004) as an indicator for light rail demand. Although light rail demand might depend on different factors than heavy rail, the number of renters does link to a group which usually has a lower income than average and thus is more inclined to use public transport. According to this paper "Renters tend to be disproportionately poor, young, located in denser multifamily housing, which may lack parking". However this factor was mainly included due to a lack of better socio-economic measures in the available data.

Number of students can also be a key indicator for rail travel. As car ownership and income among students is usually lower than the national average this group is inclined to use public transportation more often. Besides since the introduction of free public transportation for college students in 1991 in the Netherlands this group forms a large portion of the daily train users. Linked to the number of students, a higher educational institute in the vicinity of a station might also be a good indicator as this is a main destination (Wardman, et al., 2007).

Network dependent factors

The variables described here are all related to the service level provided and the relative position in the broader public transportation network. Certain features of the station and its place in the network can affect ridership in quite a strong way.

Kuby et al. (2004) included the variable normalized accessibility (or centrality) within the network as an indicator. This variable would be determined by average travel times to other stations in the network. Average travel time (including transfer time) was computed weighting all stations equally. This variable was included in contrast to the variable "distance to central business district. It was considered this distance to CBD was no long valid in polycentric cities of today.

Service frequency is first of all one of the most profound indicators of service level. A large limitation of this variable is the fact that problems might occur due to multicollinearity in-between independent variables (Taylor & Fink, 2003). One could argue for example that a higher service frequency will result in a higher ridership demand in this case. However it also can be the other way around: A higher demand for transport resulted in a higher service frequency. This is something to take into account when performing regression analysis.

Secondly, passengers find reliability and lateness of trains important. If the reliability of the lines is not as high as they expect it does reduce the perceived service level significantly. However it appeared that a high level of lateness of trains did not always deter people of taking the train (Batley, et al., 2011). The service level of the feeder modes can also be included in variables. For cyclists the presence of a bike storage facility is important while for car users a park and ride facility is more convenient. These facilities can all be included in a model as was done before a study of Brinckerhoff (1996).

In a Dutch context cycling is a relatively important feeder mode for train travel. 25% of all access trips to a mode of public transportation are made by bike. For train only this percentage rises to 29.3 percent (Martens, 2007). It was reported that passengers are not willing to travel as far for a bus stop with a lower level of service as they would for high quality public transportation (van der Bij, et al., 2010). For high quality public transport the maximum sphere of influence was about 800 metres for pedestrians and 2350 metres for cyclists. Previous research based on train station derived values of 1100 metres for pedestrians and 2600 metres for cyclists. Public transportation as a feeder mode to train stations was estimated to have an average travel distance of 7200 metres (Keijer & Rietveld, 2000).

It also matters how many destinations are reachable from a station and how often the train goes there and how well people are able to access the station. People are willing to travel further to a station which offers a better quality of service. This might result in a lower amount of people which are going to use a new station than what could be expected on the basis of a demand forecast.

Revealed preference data from the Netherlands also unveiled that 47% of all train travellers were not using the nearest train station available (Debrezion, et al., 2009). This indicates that using distance as the only indicator of travel demand has some serious limitations. Instead of using distance as main explanatory variable, Debrezion et al. suggested using the rail service quality index as main indicator instead. This indicator takes into account the position of the station within the network and the service quality provided in relation to competing stations.

Then there are certain variables describing the type of station. If the station is near a ferry or airport a variable could be included to take this into account. These kinds of stations usually receive more passengers than one would expect as ferries and planes bring in people from outside of the catchment area. Therefore a rather big error could arise between the forecasted and actual passenger demand if the variable would not be included. Finally a variable could be included to deal with terminal stations. These stations have a larger catchment area as people who live at the end of the line are willing to travel further in order to travel by train (O'Sullivan & Morral, 1996). Usually this variable is inserted as a binary variable in the regression analysis but it is the question this is the right way to tackle this problem or other modelling techniques would be needed.

Geographic dependency of variables

The effect of different factors is also dependant on the region where they are measured. Of course cultural differences between countries can be the cause of the fact that certain variables add more explanatory value to a model in one country than in another. As the U.S. is a more car centric society, one can expect that variables related to accessibility for cyclists to station areas are less of influence in rail demand in the U.S. than it would be in the Netherlands or Denmark.

However also within the same geographical region there can be differences in the explanatory value of variables in a model. As studies from Blainey (2009), Blainey & Mulley (2013) and Cardozo et al. (2014) proved that the explanatory power of variables such as number of lines, suburban bus stations, train frequency and availability of car parking all can vary across regions. Especially the difference between urban and suburban or rural areas can make a big difference and although these studies were performed in Parts of Australia, South Wales and the urban region of Madrid, Spain it can be expected that this will be similar in the Netherlands.

Conclusion

In table 3 below the most important factors in estimating rail demand found in literature can be found including the study the variable was used in. It can be concluded that many factors are thought to be able to affect rail ridership.

However, not all of these variables are suitable in a Dutch context. Whereas in the U.S. and Australia for example the mono-centric city is still quite prevalent, in a Dutch context inclusion of the variable distance to CBD would not make sense. In the Dutch situation cities are generally smaller and, especially in the Randstad area, the cityscape could better be seen as a polycentric city where trips are not as much focused on one single destination.

Other variables might become more suitable in a Dutch context such as cycling related variables. Because of the high rate of cyclists in the Netherlands, cycling accessibility could be an important variable in explaining rail ridership.

Table 3: Overview of all variables linked to ridership generation

Category	Variable	Source	Expected sign
Built environment			
<i>Density</i>	Population density	(Cervero & Knockelman, 1997)	+
	Total number of dwellings	(Blainey & Mulley, 2013)	+
<i>Design</i>	Education institutes		+
	Healthcare institutes		+
	Crowd attracting activities	(Carpio-Pinedo, 2014)	+
	Basic need services	(Carpio-Pinedo, 2014)	+
	Occasional need services	(Carpio-Pinedo, 2014)	+
	Number of restaurants and bars	(Carpio-Pinedo, 2014)	+
	Job density	(Brinckerhoff, 1996)	+
	Station area diversity	(Cervero & Knockelman, 1997)	..
	Street density (walkability)	(Gutiérrez, et al., 2011)	+
	Park & Ride	(Cervero, 2006)	+
<i>Design</i>	Parking spaces availability	(Cervero, 2006)	+
	Bicycle parking	(Kuby, et al., 2004)	+
	Guarded bicycle parking	(La Paix Puella & Geurs, 2015)	+
	Overall station quality	(Cascetta & Cartení, 2014)	+
	Architectural/aesthetic quality	(Cascetta & Cartení, 2014)	+
	cleanliness	(Cascetta & Cartení, 2014)	+
	lighting	(Cascetta & Cartení, 2014)	+
	Station security	(Cascetta & Cartení, 2014)	+
	Information availability	(Cascetta & Cartení, 2014)	+
	Climate control	(Cascetta & Cartení, 2014)	+
	Station area design	(Cervero & Knockelman, 1997)	+
	% of renters within walking distance	(Kuby, et al., 2004)	+
	Average income	(Blainey & Mulley, 2013)	-
	Number of Students	(Wardman, et al., 2007)	+
Socio-Economic	Car Ownership	(Wardman, et al., 2007)	-
	% of age of 65+	(Blainey & Mulley, 2013)	+
	% of age below 19	(Blainey & Mulley, 2013)	+
	Average household size	(Blainey & Mulley, 2013)	+
	Bus feeders	(O'Sullivan & Morral, 1996)	+
Network	Service quality	(Brinckerhoff, 1996)	+
	Centrality within the network	(Blainey & Mulley, 2013)	+
	Terminal station	(Blainey & Mulley, 2013)	+
	Distance to CBD	(Brinckerhoff, 1996)	-
	Distance to nearest IC station	(Blainey, 2010)	+
	Station Serving Airport	(Kuby, et al., 2004)	+
	Border station location	(Kuby, et al., 2004)	+
	Train frequencies	(Walters & Cervero, 2003)	+
	Station near Ferry	(Blainey, 2010)	+
	Nearest large city	(Blainey, 2010)	+

Some other variables are more kind of makeshift solutions as other suitable data was not available at the time of study (see for example the % of renters in walking distance). Later on in the methodology section it is explained which variables therefore will be included and which ones are not.

This chapter now also brings the answer on research question 1: *Which factors are playing part in the daily number of passengers using a local train station?*

Factors which are playing a part are identified from literature in table 2 and can be roughly divided into built environment, socio-economic and, network & station variables. Although this is by far a complete list it already gives an idea of the number of factors which can have an influence. However the most important variables are present in this list and although many other variables might have an influence it can be expected that most other unidentified variables will only have a minor influence on rail demand.

Secondly the geographic location of the station is of influence in the way these variables can explain travel demand. In some areas certain variables become more important than other in explaining demand and therefore the location of the station itself can also be identified as a factor of importance.

2.2 EFFECTS OF OPENING A NEW STATION

Opening a new train station will have multiple effects on the rail accessibility, total demand and personal passenger travel patterns. Opening a new station along an existing line will cause an additional two to three minutes travel time for existing passengers not using the new station. Although this does not seem much it might be just enough for certain passengers to leave the train and choose another mode in the future (Givoni & Rietveld, 2014).

On the other hand, another group of passengers will profit from shorter travel times as the new station is closer from their point of origin as the existing station. This will result in a shorter journey for existing passengers and possibly the attraction of new passengers who wouldn't travel by train in the old situation. It is especially this last group of new passengers which can make a new station feasible.

Secondly, passengers who were already traveling by train using another station might now choose to travel via the new station. Demand of other nearby railway stations might therefore decrease. This is called abstraction of demand. Depending on the service quality, frequency and accessibility of the new station, passengers will choose their new station of preference. A large share of existing rail passengers will therefore choose to use the new station. This demand abstraction and station choice is also described in recent literature (Blainey, 2010).

In the research of Blainey (2010) for example, demand abstraction is described with a multinomial station choice model. The difference between a model run with and without the new stations was then ascribed to the inclusion of the new station.

Besides demand abstraction alone there is also another effect. Although the utility of a fraction of the passengers now choosing for this new station might have been improved, the overall societal costs might have been raised considerably (Givoni & Rietveld, 2014). From forecasting passenger demand the station might have looked economically viable, however due to the abstraction of passengers this would not have been the case.

Conclusion

The conclusion on the effects of opening a new station brings back sub question 2: *"what is the effect of a new train station on station choice and mode choice"*? There are multiple effects that have to be taken into account (see figure 3). Therefore the passengers' effect of opening a new station is not always economically viable.

A new station increases accessibility onto the rail network and therefore people who have originating or destination trip in the station area are therefore getting an increased utility to use the train. This might result in an increased demand to travel by train. For some existing rail passengers the station might offer a better rail accessibility as well as the new station closer to their point of departure resulting in a change in departure station choice. Finally, a new station also causes for an extra stop on existing lines and therefore a longer travel time. Existing passengers not using the new station but are using the line will experience a longer in-vehicle time and their utility to use the train decreases slightly. This can result in a decrease in rail demand.

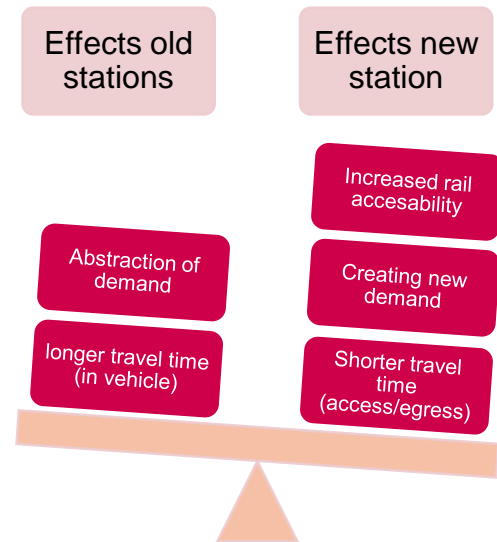


Figure 3: The balance of a new station

In contrary to many other rail estimation models such as the PINO model used by Dutch railways, the effects as stated in figure 4 should be included in the model as well. The demand model should therefore not only estimate demand on the basis of its direct environment but will incorporate the competition from other stations and network effects as well.

2.3 MODELLING NEW STATIONS

The previous sections described what factors are contributing to rail demand, what the effects of opening a new station could be, and explained the scope of which types of stations will be included in the model. This section provides an overview of multiple Ridership modelling methods which use the information from previous sections in order to make new demand forecasts.

Although there is no right or wrong model choice, each model does have its own characteristics. Each model and accompanying methodology has its strong and weak points and will be suitable in certain conditions with a certain goal in mind. Selection of the most suitable model is therefore of utmost importance.

Traditional models using the 4-step method are widely used in transport planning. These models often offer a good modelling solution on a regional scale. However there are drawbacks when the goal is only to model rail demand of local stations. The (regional based) resolution of the 4-step demand models is usually not suitable to pick up minor land use changes in the individual station areas therefore ignoring the effect of land use change on rail passenger demand. Besides, 4-step models tend to need a lot of input data which might not always be available or is expensive to gather. All together this makes 4-step modelling not that suitable for modelling the relative small areas around new proposed stations (McNally, 2008).

An alternative is found in direct demand models. Usually based on multiple regression analyses, these kinds of models are able to estimate ridership of a station as a function of station environment and transit services features (Gutiérrez, et al., 2011). However also within the field of direct ridership modelling there multiple methods to get to a final ridership estimation. Some methods are more advanced than others and therefore require more effort to produce the results. However the result might often be significantly better.

As for modelling demand abstraction and stations choice, multinomial and/or nested logit models are a better alternative as these models can model disaggregated choices of individuals. These types of models are already shortly touched upon in the previous section, a more detailed explanation is found in this section as well.

Multiple regression Models

Regression models are relative easy models to estimate and to understand, but they can be made as extensive as needed. A linear regression model could have the form:

$$Y_i = \beta_0 + \sum_k \beta_k \beta_{ik} + \varepsilon_i$$

With parameters:	Y_i	the total number of predicted passengers
	β_0	The constant or intercept
	β_k	Estimated parameter for variable k
	β_{ik}	variable value i for variable k
	ε_i	error term for variable i

However, the variables that are included can be weighted, measures and defined in multiple ways just as the cases/observations that are used. Therefore multiple methods are described including their advantages and disadvantages.

Reference Class forecasting

Since a problem of regression analysis is that the type of cases are not always entirely equal. Some groups of stations are more sensitive to certain variables as other groups. This could result in a biased forecast due to the nature of the sample group.

As encouraged by Flyvbjerg et al. (2005) reference class forecasting would prevent a biased demand forecast. This way, better estimates would be produced as for every new project the transport planner would have to look at similar projects which are already completed from the so called reference class (Flyvbjerg, et al., 2005).

Problem with this type of forecasting is that very distinct types of classes are needed. However in practice it is often hard to categorize all stations into distinctive groups. Every station is unique in the sense that the local variation of the station area is different for every station and so is the amount of passenger that will use it. If only one variable would be different at a station which is for all other variables exactly the same there is still a big chance the demand of passengers will differ significantly. And if distinctive classes can be distinguished the question remains in enough cases are available in each group.

However van Hagen and de Bruijn (2002) defined 6 station types which would be distinctively different from each other on the basis of position in urban landscape, accessibility and modal access/egress choice. Therefore within such a categorisation reference class forecasting can be a useful tool.

Euclidean distance models

Euclidean distance regression modelling is demand forecasting based on a predefined circular area around the station defined as the catchment area. With the station as centre point in the circle this type of model retrieves the number of potential passengers on the basis of number of people living or working in the catchment area. Also other variables can be included if this variable is likely to affect the passenger demand. This type of regression modelling is often used in literature as it is easy to use and understand.

In many research projects (Zhao, et al., 2013), (Liu, et al., 2013) usually a threshold of about half a mile or a series of thresholds (e.g. 500, 1000, 1500 metres) would be used to take variables as number of inhabitants or jobs in the station area into account. This is called the all or nothing approach as one is opting for a 1000 metre threshold; everyone within this threshold is attained with the same likelihood to take the train no matter this person lives right next to the station or exactly 1000 meters away.

Network Distance models

Instead of using Euclidean distances a better solution is to use the real travel distance to a station. This is relatively easily done in GIS and has already been applied in various research projects (Upchurch, et al., 2004), (O'Neill, et al., 1992), (Horner & Murray, 2004). This resolves the the problem of possible barriers (e.g. river, highway or railway track itself)enlarging the actual travel distance to the station in contrary of what could be expected when only looking at the crow-flight distance. A notable difference in ridership estimation between the two methods could be seen in the study of Gutiérrez et al. (Gutiérrez & García-Palomares, 2008) where the R^2 of a model using network distances was 0.724 compared to only 0.707 for the model using fixed distances. This indicates there the model could be improved considerable by using real network distances.

This method makes sense when features such as rivers, highways or the railway line itself forms a barrier with limited amount of bridges, overpasses and/or crossings. In such a situation the difference between a network distance model and a Euclidean distance model can grow considerably large.

Distance decay modelling

In almost all papers described above, despite of using the network distance, often fixed distances were used in order to determine the ridership. This means that there is no or little differentiation between the distance from the station and expected ridership.

In reality however this is not the case as many ridership indicators tend to lose importance when distance to the station becomes larger. Research from the Netherlands for example proved that *“people living in the ring between 500 to 1000 meters from a railway station is about 20% lower than of people living at most 500 meters away from railway stations”* (Keijer & Rietveld, 2000).

One of the first studies that took this issue of distance decay into account for transport demand modelling was the study of (Gutiérrez, et al., 2011). The number of people traveling by train for example has 10 regression functions, one for each zone around the station. This way the gradual reduction of the chance of someone choosing the train as a transport mode is modelled. However, *“in order to calibrate distance-decay functions, spatially disaggregated data on public transport use are needed”* (Gutiérrez, et al., 2011).

Demand modelling in Dutch practice

PINO (in Dutch: Prognose model In- en uitstappers Nieuw te Openen station) is the model used by the Dutch railways to make a forecast of the demand at a new station. It is a regression based model but it does include some additional features in order to improve the forecasts. It is supposed to be used for demand estimation for class 4, 5 or 6 stations. These are the smaller stations served by local trains without a node function.

The regression model is estimation a number of trips originating (home-end) and attracted (activity-end) by the new station. This done based on circular areas around the station. The circle thresholds lay at 500, 1000, 1500, 2000, 2500 and 5000 metres around the station. It is assumed that as distance from the station increases the amount of people using the station will become smaller. Therefore there is some sort of distance decay incorporated in the model.

Variables which are being used to estimate the daily use of the stations include the total population, number of jobs in the area, number of students, amount of feeders, and a competition factor because of other modes (NS, ProRail, 2006).

Geo-weighted regression methods

A relatively new development in transportation demand forecasting is geo-weighted regression (GWR). Although it was applied in other areas of study before, it is not yet that often used in transportation studies.

Problem with regular regression methods (distance decay, network distance and Euclidean distance models) is that these models are based on a set of measurements of the whole study area. From all these measurements only one regression formula will be calculated. However it is well known that certain variables will have more effect on passenger demand on one location compared with another location. It is for example plausible that the variable ‘number of regional bus lines’ is more explanatory for rail demand in rural areas than it is in the centre of Amsterdam. In Amsterdam the explanatory value of regional bus lines is mainly replaced by metro, tram and city bus lines instead. GWR therefore generates a multitude of regression formulas and the outcomes of the measurements (one for each station in the dataset) will then be interpolated.

Where a regular linear regression model could have the form:

$$Y_i = \beta_0 + \sum_k \beta_k \beta_{ik} + \varepsilon_i$$

With parameters:

Y_i	the total number of predicted passengers
β_0	The constant or intercept
β_k	Estimated parameter for variable k
β_{ik}	variable value i for variable k
ε_i	error term for variable i

A geo weighted regression (GWR) with adjusting coordinates for the dependent variable could be rewritten with $(x_i y_i)$ indicating the geographic location of the regression formula:

$$Y_i(x_i y_i) = \beta_0(x_i y_i) + \sum_k \beta_k(x_i y_i) \beta_{ik} + \varepsilon_i$$

With $(x_i y_i)$ as the location specific term. This location specific term means that the coefficients and constants/intercept are only valid for this point in space. As the GWR model allows for variation in the constants, the constants are calculated separately for each case.

This model was taken from a research on the Sydney regional rail (Blainey & Mulley, 2013). Application of this method in this instance did only saw a slight improvement of the model fit (Blainey, 2010). However, it was mentioned that this method would take "into account the possibility that parameters may not be constant across different points in space".

However, it is important to include enough cases in the geo-weighted calibration and these cases need to be distributed across the country in such a way that no region has a larger weight compared to the other regions. A combination of reference forecasting and geo-weighted regression is therefore not recommended. Applying both methods at the same time will most likely result in too few cases for the GWR in order to produce reliable results.

Demand built-up over time

With regular demand modelling usually an optimum of passengers is calculated on the basis of variables having a single point in time. However, before this optimum is actually achieved it might take several years although in research of Blainey and Preston (2009) no such evidence could be found. After usage growth rates at the new stations were compared to area mean growth no relation could be proven. But in other research it was found out this process could take up to five years (Preston & Dargay, 2005).

Reason for this build up is because people, once they developed their pattern of traveling around, are not inclined to change this pattern. This is due to the fact that people do not tend to break their habits and they often lack the information that the same journey made by rail might be more beneficial for them. There is a trade-off of opening station near new construction projects: open a station right at the start of construction with a considerable financial loss for the first few years or open the station when construction is finished but risk the fact that people are already stuck in their travel patterns.

Secondly, demand can also change over time due to external variable changes. Changes in the network elsewhere (i.e. introducing new services, closing/opening new stations), cheaper or more expensive petrol prices and changing toll rates all contributed to changing demand levels (Doi & Allen, 1986). Because of these external circumstances the effect of a new station becomes less clear due to interference with these external changes of demand.

Other limitations

A large limitation of regression analysis is the fact that problems might occur due to multicollinearity between independent variables (Taylor & Fink, 2003). If for example the variable service frequency is taken into account one could argue that a higher service frequency will result in a higher ridership demand in this case. However it also can be the other way around: A higher demand for transport results in a higher service frequency. This is something to take into account when performing regression analysis.

Secondly the availability of data can be an issue. Even if all data needed is available, it is often already outdated. Sometimes the data of the desired year is not available and the only option is to work with datasets from different year what could bring lead to some errors into the results. Therefore the quality and applicability of the resulting model is not always as good as what was aimed for.

Third limitation is that regression can only consider factors within the predefined catchment area. Passengers using the station coming from outside the catchment area are not considered in the regression. Result is that especially on transfer/ multi-modal stations the difference between predicted and actual travel demand can be rather large. Inclusion of variables such as the number of feeder lines can only partly resolve this problem.

Station Choice Modelling

Station choice modelling is suitable for determining demand changes as a result of opening the new stations and to deal with competition between stations and other modes. When a new station is opened this station is abstracting demand from existing stations. In this section therefore a description on how competition between stations can be modelled and how intermodal competition can be taken into account.

Where in general regression based modelling can be quite accurate when one is forecasting demand at a new station which is projected a considerable distance away from existing stations. This modelling technique is less useful when other existing stations are relatively close to the new proposed station as effects such as competition between stations cannot be taken into account with regression analysis. Alternative ways of modelling are therefore required.

Research in the Amsterdam area showed that a large portion of the passengers do not use their nearest train station as the access station onto the rail network. Passengers might prefer another station with a higher service level instead. A station which might be closer by the passenger's initial point of departure but with a lower level of service quality becomes less desirable (Givoni & Rietveld, 2014). Competition between stations is therefore a factor which should be taken into account. Therefore, in order to prevent large errors in the demand forecast other modelling techniques might be better suited for demand forecasting in areas where the existing station density is larger.

In research from 2004 two different logit models were tested when modelling station choice and access to rail network. The first model tested was a conventional MNL model. However, *'it was found that abstraction from competing stations took no account of their proximity to the origin station, and this was obviously a limitation'* (Lythgoe, et al., 2004). The second model was a cross-nested logit model. This model resolved this issue and had a better fit than the conventional MNL model. However, a big limitation in this research was that all access trips to the station are considered being done by car. In a Dutch case this would be far from realistic.

In research of Givoni and Rietveld (2014) it was calculated what the effect would be upon closing or opening new station in the greater Amsterdam region. By again using a nested logit model the utility of various access modes and station was calculated. This was done twice in order to compare the situation before and after closing or opening of a new station on an existing line. The difference in utility can then be interpreted as the benefit/loss of opening or closing a station.

Adding a new station would result in a slower travel time for existing passengers reducing their utility of using that line, but on the other hand it increases utility for using the line for people living and working close to the new station as it increases their utility of using that station. Closing a station would have the same affects but in this case reversed. Other passengers not using the station would enjoy a faster travel time, but passengers who were using the station would suffer longer travel time as they would need to travel to the next best station according to their utility function (Givoni & Rietveld, 2014).

However this effect of closing or opening a station was expressed in way which is rather hard to understand for non-experts. Closing one of the stations in this study would cause an increase of the log sum with 419 "disutility units". This can, according to the study be translated into an average of 2.18 euros of loss per rail departure for every passenger who was using the station with the use of a value of time of 10 euros per hour.

In another study which focussed on calculating the competition effects between two stations the changes before and after opening a new station were simply mapped. These changes consisted of the difference of the probability that a postcode area would use a certain station (Blainey & Evans, 2011). As this would be mapped before and after the introduction of a new station, it made it insightful of what the effect would be on station choice. However these probability differences were not recalculated into actual loss of number of passengers in this paper.

As railway station choice is thought to be dependent on multiple variables such as their accessibility, distance from point of departure and level of service. As the combination of these factors plus the access mode determines which station is chosen in the end, there is a need for a way to model this choice behaviour. Debrezion et al. (2009) introduced a so called rail service quality index. This index categorised stations on the basis of four different indicators:

1. Train frequency: As a high frequency implies shorter average waiting times passengers should prefer a station with a high frequency service.
2. Network connection: How well is the station in question connected with the rest of the network? This can be estimated by calculating the total number of destination one can reach without a change.
3. Service level: A passenger usually prefers a station with the highest service level. This means they prefer trains going from departure to destination as quickly as possible. Intercity train stations are therefore preferred above sprinter train stations.
4. Monetary costs: The higher the costs are for a train ticket the likelier it is they seek an alternative route or mode.

Based on this indicator a double constrained spatial interaction model was built which was the basis of their further analysis with the use of a multi-logit choice model. The RSQI therefore formed one of the main variables in the multi-logit model together with the access mode related variables.

Debrezion et al.(2009) then used similar nested multinomial logit models as also was demonstrated in the paper of Givoni and Rietveld (2014) in order to model station access mode and station choice. It was assumed that the choice of access mode and station are made simultaneously. There were four alternative nests in total: walking, cycling, car and public transport. They used a nested model in order *"to deal with the independence of irrelevant alternatives assumption of the standard multinomial logit model"*.

As no data on individual passengers was available the utility per mode was calculated for each zip code area. The variable "car ownership" was used in order to determine the access mode. A high car ownership in a zip code area would result in a higher utility for using the car as an access mode and decrease the probability that bike or walking would be used.

In the lowest level of the choice tree the utility of the three nearest stations would then be calculated. The formula used to estimate the utility functions would include variables such as the presence of a bicycle parking facility or P+R facility. If this would be the case the utility of, in this case bike and car would be increased.

Main conclusion of modelling station choice is that utility theory with discrete choice modelling is often used in combination with the corresponding probability a passenger from a zone is choosing a station. The use of utility theory gives the opportunity to also include factors that determine the attractiveness of certain stations such the inclusion of variables such as the availability of bike parking, car parking or other services. The rail service index which is calculated for every station is a good example for this.

Feeders and intermodal competition

Besides competition between stations, there is also competition between modes. Especially in urban areas where alternatives such as metro, tram, and bus are present, rail travel can suffer some losses because of people using these alternatives. A good connectivity between these other modes and the new station can also result in these other modes acting like a feeder network causing the new station to receive more passengers on a daily basis than what can be expected based on a regression analysis.

Whether these other modes will act like feeders or competing modes is depending on the direction, destinations, and speed of these lines. In order to model feeders and intermodal competition, also other modes of transport besides the train should be taken into account by enlarging the scope of the model. However in all previous models touched upon, it was assumed that it was already decided to only use rail based trip to calibrate the model on. Modelling competition thus requires the mode choice to be modelled as well. Hence why all intermodal stations were removed from the analysis in the research of Blainey & Mulley (2013), a regression analysis to estimate demand of train stations in the Sydney area.

Using a zonal gravity based models to calculate the number of trips in an origin/destination matrix the factor mode choice can be incorporated. A multi-logit choice model incorporating mode and station choice was then used to make a demand estimation of station usage (Wardman & Lythgoe, 2004) and (Wardman, et al., 2007). It should be noted these researches were based on rail tickets sales data, something which is not available for this thesis, and was focussed on rail journeys longer than 40 or 80 kilometres whereas this is not always a realistic threshold for the Dutch railway system.

This large threshold value was deemed necessary in order to make a distinction between access modes and the total travel distance. Therefore only trips longer than 40 or 80 kilometres were taken into account. Also the mode choice consisted of choosing train or other mode without elaborating what the other modes could be used (e.g. bus, car, and metro). This "other mode's" utility function was solely based on the costs of traveling along the road network. No timetable information on any public transportation alternatives had been included.

Another study of Blainey and Preston (2009) did take the possibility for different mode choices into account. In the study also the bus was considered as a modal choice. Using a direct demand model, the total number of trips from each zone to another was estimated. Also the modal split of these trips (bus and train) was calculated. However due to a lack of timetabling information on bus travel times and insufficient results, the final model only contained the generalised costs by traveling by car.

Conclusion

Regression analysis is a suitable tool for estimating the total ridership of a new station. By weighting density variables (such as population, number of jobs) with the use of distance decay, and using network instead of Euclidian distance, enhanced variables can be made. These enhanced variables can then be used in the regression analysis for improved results.

For effects such as demand abstraction and mode choice changes however, logit choice models are a better alternative. However, disaggregated travel data is required for calibration of these models.

In short there are three main reasons for generating a station choice model next to a regression analysis as well:

- I. Using the distance decay weighted number of inhabitants as explaining variable for relatively isolated stations might work very well for estimating ridership. However when more train stations are located closer near each other only using distance decay might no longer be sufficient. Problem is that at some point the catchment areas of the distance decay functions will overlap each other. Taking no measures to resolve this will result in double counting the same inhabitants whereas in reality people can only choose one station for a trip.

In order to resolve this problem, Thiessen polygons are commonly used. This way every inhabitant will simply be assigned to their nearest station. This however can be realistic when all stations offer the same service level and same type of facilities. However in reality the service level and facilities available at each station differs which causes a preference for certain station types above others.
- II. Current models are static in such a way that the addition of a new station will not have an effect on the other stations. They do not give any information on how many new passengers a station can generate and what part of the passengers using the new station are abstracted from existing stations. This however can be an important factor in the decision to open a new station or not.
- III. It is known that access mode choice and station choice are influenced. A cyclist might choose a station with good bicycle facilities while a car driver will need good parking facilities. It can also be the other way around that access mode choice is determined by how good the facilities are for a certain mode.

To answer sub question 3: "*which methods are available in order to estimate the daily number of passengers of a train station*"? It can be said two types of separate modelling types can be recognised:

The first categories of rail demand models are so called direct demand models. Based on variables of the station, socio-economic factors, population, and job factors demand is calculated. Demand is therefore a function of certain variables of the station and station area. These models are therefore also aggregated models as no personal trip information is required to use this kind of modelling. This type of modal is especially suitable relative simple way to estimate the demand of a new station.

Second type of modelling is closer to traditional traffic modelling and does contain at least some if not all steps of the four step model. Therefore this type of model can take into account mode choice, station choice, travel times and congestion levels depending on how advanced the model is. This type of model is more suitable for research into additional effects of new stations such as demand abstraction, competing modes, modal shift and the amount of new rail passengers as opposed to existing users.

2.4 STATIONS IN THE DUTCH PRACTICE

Based on the variables and factors explained in the previous sections, stations can be divided into several categories. A main indicator for categorizing stations is often the service level. The model developed in this thesis is aimed for ridership estimation for sprinter train stations. In Dutch practice a sprinter train station is exclusively served by sprinter train services. These sprinter train service is a train service which usually stops at every station along the line. The service quality of these stations is therefore lower compared to the intercity train stations. This latter station type is also served by the faster intercity trains which only stop at stations in the larger cities.

However, there are exceptions. Certain local train stations are served by intercity trains on some parts of the day such as station Amersfoort-Schothorst. Also quite common is that intercity trains act as sprinter trains on the final part of the line such as certain train series on the line Zwolle-Leeuwarden and Zwolle-Groningen. Another definition for sprinter train stations is not defined by service quality but by catchment area. This way sprinter stations could be seen as "stations serving local transport needs" (Preston, 1987).

In some cases sprinter stations are also referred to as (sub)-urban stations or commuter stations. This is also not the correct term as using this term would imply that only stations used for commuting or that only new stations in urban areas would be taken into account. In this thesis the goal is to take every new sprinter station into account and thus also stations in rural areas which are usually not covered within the definition of 'urban' or 'commuter' stations.

In the Dutch document "Typisch NS: Elk station zijn eigen rol" (2002) Dutch station were even further categorised into 6 types of stations:

1. A large station in city centre of large city.
2. A large station in city centre of middle-sized town.
3. A suburban/parkway station near a bigger city with node function.
4. A station near centre of small town.
5. A Suburban/parkway station without node function.
6. A station near small village/town.

In a Dutch context it means that in practice only category 1 and 2 are served by intercity trains. It is however unlikely that a new category 1 or 2 station will be opened and these stations are therefore considered beyond the scope of this thesis. Category 3 stations are incidentally served by intercity trains. Stations of category 4, 5 and 6 are in general only served by sprinter trains (Van Hagen & De Bruyn, 2002). However, there are exceptions as certain type 3 and 4 stations are being served by intercity trains on a regular basis.

In practice it can be assumed that all stations of type 3, 4, 5, and 6 receive less than 3500 daily passengers on an average weekday. This group of stations will consist of roughly 75% of all train stations in the Netherlands and are often serving only a part of a city or town. These stations therefore have a local function instead of a regional or national function. All stations opened in the last decade are currently receiving less than 3500 passenger a day on average.

Dutch railways also assigned stations with an official intercity status. This list includes stations from type 1, 2, 3, and 4. Since type 1 and 2 stations are out of the scope it is the question whether to include the type 3 and 4 intercity status stations or not. Based on the regression results it is decided whether to include these lower ranked type 3 and 4 intercity stations or not.

Proposed train stations

In the Netherlands in the current situation there are about 40 proposals of new stations and the demand estimation model resulting from this thesis should be able to make demand estimations of these stations. The progress of each of these stations varies from initial proposals to complete worked out designs which will be built within short notice. All of the proposed stations are sprinter stations which are planned to only being served by sprinter train series or intercity train series on a limited basis.

A list of proposed stations along the main railway lines and some decentralised lines in the Netherlands can be found below in appendix I (Ministry of environment and infrastructure, 2014). This list is not complete since local railway station proposals that don't need funding from the national government are not found in the list published by the ministry of environment and infrastructure.

Whether or not some other stations will be built sometimes depends on accompanying construction plans of new dwellings and office buildings (i.e. Leeuwarden Werpsterhoek). Without the additional dwellings the proposed station often will not be economically viable. Especially since the economic crisis in 2008 these stations are less likely to be developed in the foreseeable future.

Other proposed stations are depending on additional infrastructural measures in order to implement these new stations in the current timetable. Otherwise there wouldn't be enough capacity to deal with the additional dwell time caused by the extra stop the train has to make. A possible station at Staphorst for example is hard to implement within the existing timetable and infrastructure although the station is deemed feasible when it comes to the estimated numbers of passengers.

The majority of the proposed stations from appendix 1 are not feasible in the first place because of the low amount of passengers which is expected to use the station and are also not expected to become feasible in the near future. Plans for these stations are often suspended and might only be reconsidered after 2028 in case the situation has changed. After calibration, and validation of the demand estimation model, the aim is to do a demand estimation for the majority of these stations.

3. METHODOLOGY & DATA

In chapter 2 the various factors determining passenger rail demand and the effects of a new station on overall local demand were identified based on literature. In this next chapter, this knowledge is used to present a research approach and to develop a method that will achieve the goal of this thesis: *“To develop a demand forecasting method for new train stations which is able to provide a ridership estimations of new stations based on departure station choice and network accessibility”*.

In the second part of this chapter the data which is used in this thesis is presented. Because some data is not available or only available on a limited basis, some considerations have to be made on which data is to be included.

3.1 RESEARCH APPROACH

For forecasting the daily number of passengers using a station, two main methods were identified in literature which might be able to produce accurate results. However both methods have their advantages and disadvantages.

Regression modelling is a useful method for relatively ‘isolated’ local stations. These stations should have a limited catchment area, without feeding or competitive public transport lines and without competing stations in the vicinity. If this is the case the modelling results can be quite accurate and disaggregated data is not needed in order to conduct use this method. However in the Dutch context this is not often the case, especially not in the Randstad area.

Opposed to regression modelling there are more traditional modelling methods which include station choice and modal choice. These methods however take more time and are less sensitive for local variation in land use or other local factors. However they do take into account feeding modes and station choice based on utility functions. Therefore in a more complex station environment this modelling approach is more suitable. Also this type of modelling gives the opportunity to produce an insight in demand abstraction and changes in station and mode choice. However this type of model also needs disaggregated data input.

For this research is has been chosen to:

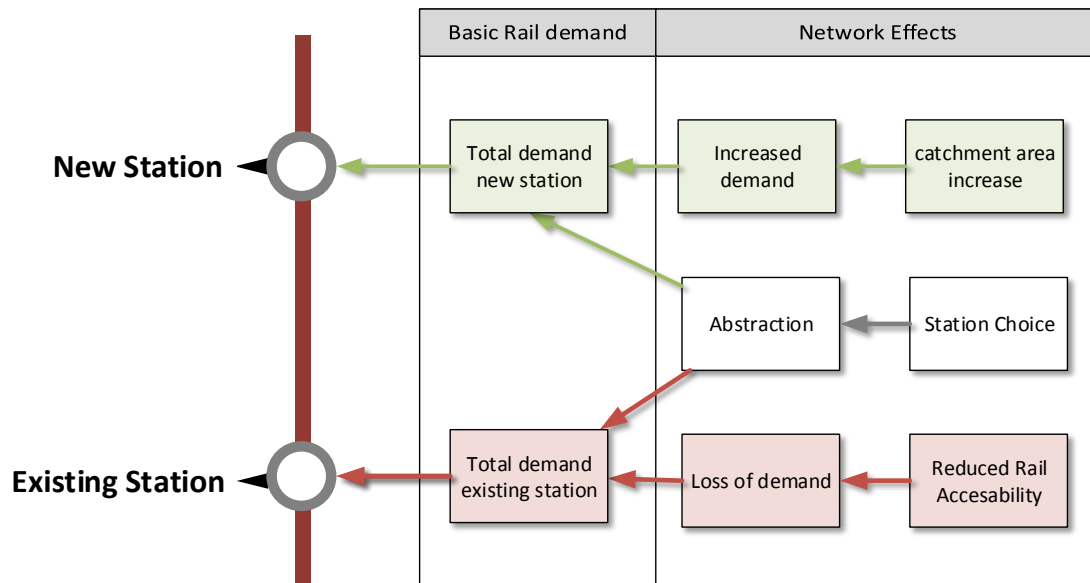
- A choice model based method will be used to estimate a model based on disaggregated trip data. This modelling method gives the opportunity to also research the effects on station choice locally and allows the enhanced population and other density variables to be weighted according to the station choice before entering the regression analysis.
- Use a regression based method on a national level. This model is aimed at making ridership forecasts and should be applicable in the whole country. Input data will consist of variables measuring the density, accessibility and quality of the individual stations.

After these modelling steps, an overall assessment can be made of a new station and its effect on the local transport system. This way it can be determined whether or not the new proposed station might be feasible or not on the basis of passenger flow and demand. A total overview of the whole project in the form of a conceptual model is depicted in figure 1. Squares depict steps in the overall process whereas ovals depict the necessary input data for these steps.

3.2 ANALYTICAL FRAMEWORK

The effects of the opening of a new station are depicted in Figure 4 (based on figure 3). There are three main effects: increased demand as a result of a larger catchment area, shift in demand (on station level) due to changed preferences and, loss of demand due to increased travel time.

Figure 4: The effects that can be expected when adding a new station



Increased demand

Demand for rail transportation can be increased by enlarging the catchment area of the railway network. This is done by opening additional access points, or stations. By opening a new station it is therefore assumed that more people are gaining access to a railway station within an acceptable distance. This will increase the demand for rail transport. The first model step will therefore be to define the catchment area of a station and the resulting total demand.

Demand abstraction

Besides attracting new passengers, also existing passengers are abstracted from other stations. This is called demand abstraction. Since many new railway stations are close to existing stations, demand abstraction is common. On an overall scope this will not affect total demand for rail transport. However on a station level this effect can cause a significant reduction in demand at other stations. Too much demand abstraction is therefore not desirable unless the goal of the new station is to divert passenger flows. This demand abstraction can be modelled with the use of a station choice model since this effect only redistributes existing demand over the stations based on station specific characteristics.

Loss of demand

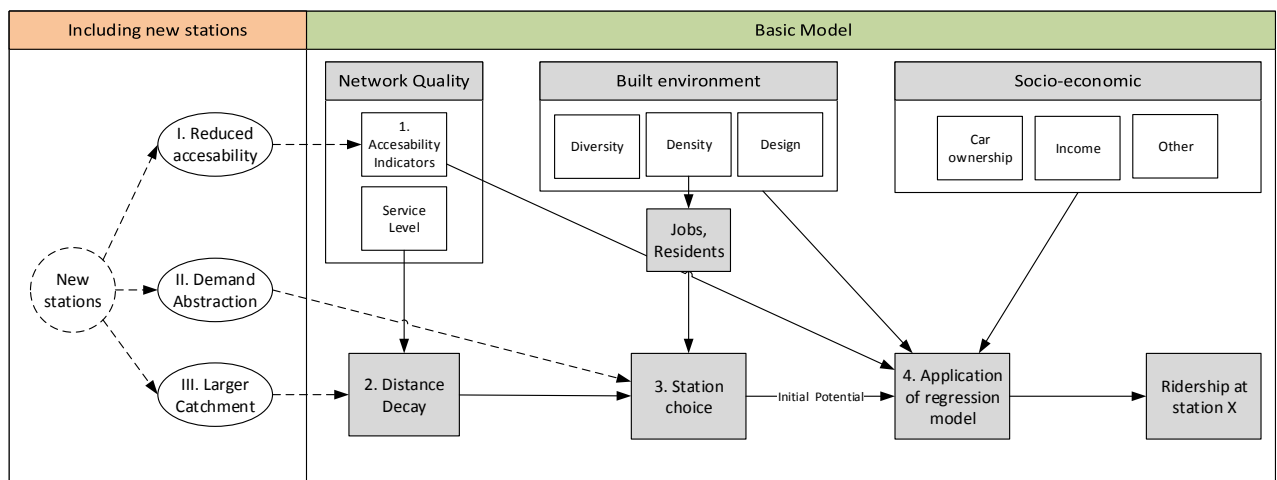
A loss of demand as a result of opening a new station is caused by the reduced network efficiency. More stations means an increased average travel time across the entire railway line. Loss of demand can thus be affecting a large number of stations. The size of this effect depends on the flow of travellers passing this new station. This effect will be captured in an accessibility indicator which is calculated with the use of Omnitrans.

A new station in the bigger picture

In the literature, three categories of variables were identified in order to estimate rail ridership: built environment, socio-economic, and network variables. As mentioned in the research approach (section 3.1), certain variables (population, jobs, students) from the variable category "density" will be enhanced/weighted according to a station choice model and distance decay weighting.

In the basic situation (Figure 5: Conceptual model for ridership estimation) it means that for example the total population within a certain distance around the station is weighted according to distance decay curves. Next step is then to apply a station choice model to assign a station to every distance decay weighted population unit (for example a postcode area). Final step will be the application of a regression model.

Figure 5: Conceptual model for ridership estimation



When a new station is added to the network the three effects that can be expected (increased demand, abstraction reduced demand) will have their effect on the basic model (dashed lines in figure 3):

- I. *Reduced accessibility* will affect the existing stations only in the final modelling step when applying the regression models. The rail accessibility variables should be estimated every time a new station is added to the dataset. More stations along the line will mean a longer general travel time. Depending on the size of the flow passing the new station for existing stations, it will decrease the rail accessibility in some degree.
- II. *Demand abstraction* will become visible when the station choice model is applied. The share of existing stations will drop in certain areas while the share of the new station will become higher.
- III. *The increased demand* as a result of a larger catchment area is estimated when the distance decay curves are applied. Since the distance to a station is reduced for many postcode areas, it is expected these areas will have a larger share of rail users then they had without the new station.

3.3 MODELLING STEPS

In total four steps are needed for a model as described in figure 3:

1. A measure for rail accessibility
2. Distance decay curves
3. A Station choice model
4. Regression models

Rail Accessibility

Omnitrans will be used in the early stage of the analysis in order to generate variables to be included in the regression and possibly the station choice model. A rail quality service index (RSQI) will be generated for each station to be researched in a similar way as has been done in the research of Debrezion and Rietveld (2009). Therefore it is needed to generate a distribution matrix of rail travel in-between all stations. On the basis of the modelled trip flow of this matrix the centrality and relative accessibility relative to all other stations in the Netherlands can be determined. As up-to-date data is not available for all stations in the Netherlands, a mix of 2010 and 2013 ridership data is used.

Distance decay curves

The catchment area of a station will be determined on the basis of distance decay functions. Using disaggregated trip data of which the point of origin is known, the probabilities of using a train station at several distance thresholds will reveal distance decay curves. Since there are strong indications, the station type (which is based on the service level) is an important indicator for the catchment area, distance decay curves will be based on station type.

Station choice model

A station choice logit model will be estimated similar to Debrezion et al. (2009). This model will then be used to derive the number of passengers changing from departure station as an effect of opening (or closing) a new station. This model will be calibrated on disaggregated trip data. This model is then applied on a six digit postcode level.

The station choice model will distribute all train travellers to a station based on the station characteristics. This is how competition between stations can be included in the model. Ultimately several enhanced density variables which are weighted by distance decay and the station choice model will be used as input for the regression (the initial potential).

Regression analysis

First of all, the dependent variables used in the regression should be a measure for rail ridership. Most common in literature as well as in daily practice is to use the total number of passengers boarding and exiting at a station on an average working day. Since most figures for ridership are available in this format (see section 3.4) this type of figure will be used as the dependent variable in the regression.

As presented in the literature review, there are various ways of doing a regression analysis. It appeared from literature that regression analysis using real distances instead of crow-flight distances resulted in a much higher explained R^2 and more realistic catchment areas as they take into account the barriers (rivers, infrastructure etc.) which might be present in the vicinity of the station environment. Therefore, real network distance will be used in this thesis in order to calculate the catchment areas.

Secondly, including the distance decay weighted density variables would also increase the explanatory power of the model. The inclusion of the station choice model will take competition between stations and station preferences into account. The ultimate population variable used in the regression will thus be a variable that takes into account the actual distance to the station, the

distance decay effect and, preferences for certain stations as defined in the station choice model. Since there are many types of stations, separate regression models will be estimated for different types of stations.

In order to enhance the exploratory power of the model the conventional global model will be calibrated using GWR (geographic weighted regression). This was demonstrated in several papers before as has been discussed in the literature review.

During this calibration process ArcGIS will be used to recalibrate the best models to achieve a better overall fit. This is done by allowing spatial variance among certain variables. The outcome might be that the whole model will fit better when spatial variance is allowed or that only a part of the model is in need of spatial variance. After the geo-calibration, the final model fit might have been improved considerably. When the variables' effect on the dependant variable may vary across regions, the regression formula will then take that into account from now on.

3.4 DATA

For all consecutive modelling steps various sources of data will be used. This section will give an overview of all data used in this thesis. Roughly five different data types can be distinguished:

- Ridership per station
- Rail network dataset
- Road Network dataset
- Disaggregated trip data
- General model variables

Ridership per station

Ridership data is essential in this thesis since the final ridership model will be calibrated and validated with the use of this dataset. Ridership per station (boarding and exit) per average working day is freely available up to the year 2014 (NS, 2014). This data contains only the stations served by Dutch Railways (NS). Ridership figures from stations served by other transportation companies are often not freely available. Figures from the Merwedelingelijn from Dordrecht to Geldermalsen, operated by Arriva, are freely available as well (Netwerk Zuidelijke Randstad, 2015). This makes a total of 300 stations in the Netherlands of which ridership figures are available up to 2014.

Rail network dataset

A rail network dataset including the corresponding properties of all links such as speed, length and, service level is necessary for making a measure for the rail accessibility. This rail accessibility will be a figure to explain how well the station in question is connected with all other stations in the network. To do so a passenger flow model is estimated in order to identify the most important destination stations as seen from the origin station. Therefore, for every origin-destination pair the distance, travel time and, number of transfers will be needed.

The basic rail model was available for use at Goudappel Coffeng based in Omnitrans traffic modelling software. This model already contained most rail links with the corresponding properties and timetable information up to 2013. A few adaptations of this data source were required in order to make this dataset fully suitable for this thesis. This was done by adding the stations opened between 2013 and 2015 to the network as well.

Road Network dataset

The Road network dataset is needed for distance calculations on the 6 digit postcode level preferably using ArcGIS. Based on these distance calculations, a distance decay weight can be assigned to a postcode. Secondly this dataset is necessary for the calibration of the station choice model since it is

expected that the distance from the origin to the departure station will be an important factor in station choice.

Since the access mode to reach the departure station can be by car, bicycle, public transport, or by foot, it is required the network that will be used is detailed enough to be able to model all of these modes. The network should therefore not only include the main roads. Minor walkways, cycling paths and pedestrian passages are important as well since this can increase a catchment area of a station significantly.

The use of the freely available open street map network (openstreetmap.org) did meet these requirements and it could be directly imported in ArcGIS. Only a few adaptations were needed to make this network suitable for use. These adaptations consisted of adding underpasses at mostly the larger stations in the dataset which were missing in some occasions.

Disaggregated trip data

Disaggregated trip data is essential for calibrating the distance decay curves and station choice model. It must contain the point of origin of a trip and the choice of the departure station. This way the distance a traveller is willing to travel for boarding a train can be derived per station type.

Freely available trip data in the Netherlands such as the MON travel survey does not contain information about the departure station. It would be possible to derive distance decay functions from this dataset but the functions cannot be established for specific station types. Only general decay functions would be possible.

This is way another source of disaggregated trip data is used. The Stedenbaan survey conducted by the University of Twente, contains almost 1500 cases of revealed preference trip data including point of origin and the choice of departure station. The survey was conducted online in 2013 in the Dutch province of Zuid-Holland. Further details about this survey will be given later on (

4.2 Distance Decay Functions).

General model variables

The general model variables are all other variables included in the ridership estimation model. They can be used as attributes in the station choice model and as independent variables in the regression analysis. All variables that will be tested for this model will be described here in the same categorisation as described in the literature review (2.1 Factors determining Basic Rail demand).

Built environment factors: Density

Density variables are identified as variables describing the density or count of attributes that directly results in rail demand. Most important density variable is the number of inhabitants. From literature it is known this variable can explain a large portion of total rail ridership. Since it is important to have this data on a detailed level, the number of inhabitants per 6 digit postcode area will be used. This dataset was published by the Dutch Bureau of statistics (CBS) on basis of data from 2013.

Density variables which are suitable for explaining ridership on the destination side of the trip are the number of jobs and total student enrolment. Data from the number of jobs was derived from traffic model zones from the national traffic model (NRM). Although the number of jobs is especially detailed around The Hague and Rotterdam, the rest of the country was represented as well, but in a less detailed level. In order to represent the data on a 6 digit postcode level, all jobs in a zone were evenly distributed to all postcode point in the same zone.

Student enrolment was available from the web portal "data.overheid.nl" which contains freely available datasets from the national government. This included the location and the number of students for every high school and all higher education up to 2014. Only problem with this dataset is that colleges with multiple locations are only assigned a total number of students over all locations. For suitable use

of this data, the students from colleges with multiple locations were evenly divided over all locations. As in reality one location might be significantly larger than another location this solution introduces an error in the data. However, it is expected that despite of this error this variable will allow for some explanation for total ridership.

There are some other density variables which are counting the number of business which can also be subdivided into certain business sectors. This data is available in a four digit postcode level from the CBS Statline. Variables and their corresponding names that are used are the total number of businesses in:

- | | |
|---|-------------|
| • Total number of registered businesses | A_BEDV |
| • Number of business in the hospitality sector (restaurants, cafés, hotels) | SOM_HORECA |
| • Number of business in the touristic sector | SOM_LEISURE |
| • Number of Shops/Retail sector | SOM_SHOPS |
| • Number of businesses in the commercial/finance sector | A_BED_FIN |

Also available from the CBS Statline on a four digit postcode level are the availability (and count) of certain services within a 3, 5 and 10 kilometres radius from the postcode zone in which the station in question is located (based on the road network). The services for which the data is available and their corresponding names are:

- | | |
|---|------------|
| • High school (VMBO) | AV#_ONDVMB |
| • High school (HAVO/VWO) | AV#_ONDHAV |
| • High school (any) | AV#_ONDVRT |
| • Cinemas | AV#_BIOS |
| • Theatres | AV#_PODIUM |
| • Hospitals | AV#_ZIEK |
| • Supermarket | AV#_SUPERM |
| • Basic need retail | AV#_DAGLMD |
| • Department stores | AV#_WARENH |
| • Attractive locations (museums, amusements parks etc.) | AV#_ATTRAC |

Finally there are some general density variables also from CBS Statline: total population density (Bev_DH) and the area address density (OAD) per four digit postcode area. For these two variables an average was taken from all zones around the station in a 5 kilometre radius corrected for the total area each zone is represented in this buffer around the station.

Built environment factors: Diversity

As a measure for diversity the land use mix (LUM) as described in the literature review will be used. As an input for this variable the total area used for residential, retail/small business, and commercial is used. This data is derived from the BBG (bestand bodemgebruik Nederland) from 2010.

The data from the BBG are also included as separate variables. In a 5 kilometre radius from all stations the total area used for infrastructure (wegverkeersterrein), residential (woon), small businesses (detail_horeca), culture (cultuur), commercial (bedrijf), parks (park), sports (sport) and, "other" is derived. The category other is undeveloped land or land in use for agricultural purposes. Therefore from each station also the percentage of the total area which is developed is derived (Opp_bebouwd).

Built environment factors: Design

The first variables in this category are variables describing the facilities present at the station itself. This includes availability of rental bikes (Bicycle_rental) and guarded bicycle parking (Bicycle_parking)

both taken from the website of Dutch railways (NS) in 2015. The availability of Park & Ride facilities is included in two ways: The total number of available parking spaces, and a measure in the size of the park and ride facility ranging from 1 to 4. 1 means 0 to 50 places, 2 is 50 to 100 places, 3 is 100 to 200 places, 4 is over 200 places. These variables are based on data from the ANWB (Dutch car-user organisation) freely available on their website. Only some smaller station on which no data was available are included manually with a count based on the use of Google Earth.

Since subjective for all stations in the Netherlands is not available and also hard to acquire, variables such as station security, cleanliness, lighting and overall station quality are not included. As for the architectural quality some variables were included. First of all the architectural style of all stations (mainly based on the year of opening) was categorised. Five architectural categories were derived:

1. No distinctive architectural style (basic station)
2. Station building from before 1945 but no longer in use
3. Station building from before 1945 and still in use
4. Station built between 1945 and 1999
5. Station built after 2000

Note: stations opened before 1945 of which the station building was rebuilt later on, are considered stations from after 1945.

Next to this categorisation a binary variable is included which is 1 if roof cover at one or more platforms is available and 0 if not (Overdekt_perron).

Socio-economic factors

Most socio-economic variables were derived from CBS Statline on a four digit postcode level. Core property of these variables is that they give additional information about the density variables as described earlier. The following socio-economic variables are included:

• Percentage of non-western immigrants	P_N_W_AL
• Average House value (WOZ)	WOZ
• Percentage of homeowners	P_KOOPW
• Percentage of empty/depilated dwellings	P_LEEGSW
• Percentage of dwellings built after the year 2000	P_WN2000
• Average number of cars owned per household	AUTO_HH
• Total number of cars per postcode area	AUTO_TOT
• Total number of company owned cars per postcode area	AUTO_BED
• Total number of cars per square kilometre	AUTO_LAN
• Average income	GEM_ink_pi
• Percentage of people aged between 0-14, 15-34, 35-65, 65-74 and, >75	P_0014 etc.
• Percentage of non-active persons (unemployed, retired)	P_NIETACT
• Percentage of household consisting of 1 person	P1P_HH
• Multiple persons and no children	M_HH_ZK
• Multiple persons household with children	M_HH_MK

Network dependent factors

Data on the number of lines and frequency on these lines for rail travel as well as bus, tram and, metro were taken from timetable data from the transport operators from the year 2013. Per station the following variables are used:

• Number of lines of bus, tram and metro combined	BTM_NOL
• Number of metro lines	metro_NOL
• Number of tram lines	tram_NOL
• City Bus lines	Stadsbus_NOL
• Regional bus lines	Streekbus_NOL
• Number of sprinter train series	sprinter_NOL
• Number of intercity train services	IC_NOL
• Frequency on lines of bus, tram and metro combined	BTM_freq
• Frequency on metro lines	metro_freq
• Frequency on tram lines	tram_freq
• City Bus Frequency	Stadsbus_freq
• Frequency on regional bus lines	Streekbus_freq
• Frequency on sprinter train series	sprinter_freq
• Frequency on intercity train services	IC_freq
• Total train frequency	Freq_Tot

On the basis of data about the reliability of passenger trains in the Netherlands (taken from rijdendetreinen.nl) the variable Delay_2013 is derived. It gives the number of disruptions of the regular service in 2013 for the station in question.

A series of binary variables is included as well. The variable "Regio_Verv" is 1 if a regional operator runs the trains and 0 when NS is the operator. The Variable "Randstad" is 1 in case the station is located in the Randstad area, 0 otherwise. If the station has an official intercity status the variable IC_service will return 1, 0 otherwise. "IC_Partial" is 1 in case some intercity trains stop at the station despite the station might not be officially given the intercity status. The binary "Terminal" is 1 in case the station is at the end of the line, 0 otherwise. The variable other_St_2013 gives the number of other stations in a 15 kilometre radius and is thus a measure for station density.

Finally also the average distances to multiple types of services are included. These variables were derived from CBS Statline as well and include the average distance to:

• High school (any)	AF_ONDVRT
• High school (VMBO)	AF_ONDVMB
• High school (HAVO and VWO)	AF_ONDHV
• Nearest highway on-ramp	AF_OPRIT
• Cinema	AF_BIOS
• Theatre	AF_PODIUM
• Nearest type 1 or 2 station	AF_OVERST
• Attraction (such as museum, amusement park etc.)	AF_ATRAC
• Department store	AF_WARENH

On the basis of these variables also an average distance was calculated. This variable based on the average distance to any high school, a cinema, a department store, a theatre and, the nearest type 1 or 2 stations. This variable (PROXIMITY) will thus be giving a measure for remoteness relative to the larger towns and cities.

Summary

To conclude, Table 4 gives an overview of all sources used for retrieve variables that serve as input for the regression analysis. A complete list of all variables used in the regression can be found in appendix 2.

Table 4: Overview of all data sources

Variable	Source
Rail network	National rail model (Goudappel Coffeng)
Road network	Open street map
Disaggregated trip data	Stedenbaan Survey (University of Twente)
Ridership per station	NS & Monitor regiospoor Zuid Holland
Population	CBS
Jobs	Abstracted from NRM, Rijkswaterstaat, 2011
Location and number of students per school/college	data.overheid.nl, 2013
Land use	BBG, 2010
Socio-economic	CBS Statline, 2014
Station Specific	NS, 2014
Data on frequency and number of lines	Operator Timetables, 2013
Services delayed/cancelled	rijdendetreinen.nl, 2014
Relative accessibility to all other rail stations (RSQI)	Generated in Omnitrans

3.5 MODEL VALIDATION

A validation of the station choice and regression models is needed before the models can be implemented into practice. This means that it will be verified that the models are able to give reliable ridership forecasts. In this research the models will be validated with use of the back casting method.

With the back casting method the number of passengers using an already opened station will be “forecasted” with data which was available before the station was opened. As the current number of passengers is known, applying the model for this station with data of before the station opened should provide some information on how the model can provide accurate forecasts.

In this case the demand for all stations opened between 2006 and 2014 will be forecasted with 2005 as the base year. The results will be evaluated in relation to the actual known ridership. In total 24 stations are included in the back casting validation dataset (Table 5).

As the model will be fed with data from 2005, all variables used in the regression analysis should be available for 2005 as well. This means that infrastructure improvements such as the opening of the Hanze line (a new railway track between Zwolle and Lelystad) should be taken into account as well. This can be adapted manually in Omnitrans. Data from 2005 is available for all other variables.

Station	Year of opening
Arnhem Zuid	2005
Den Haag Ypenburg	2005
Twello	2006
Helmond Brandevoort	2006
Amersfoort Vathorst	2006
Tiel Passewaaij	2007
Utrecht Zuilen	2007
Purmerend Weidevenne	2007
Amsterdam Holendrecht	2008
Amsterdam science park	2009
Maarheeze	2010
Sassenheim	2011
Hardinxveld Blauwe Zoom (Arriva)	2011
Slidrecht Baanhoek (Arriva)	2011
Halfweg	2012
Almere Poort	2012
Kampen Zuid	2012
Dronten	2012
Utrecht Leidsche Rijn	2013
Maastricht Noord	2013
Nijmegen Goffert	2014
Apeldoorn de Maten	2006
Hengelo Gezondheidspark	2012
Apeldoorn Osseveld	2006

Table 5: Station to be used in the validation phase

The station choice model will be validated using a similar technique as the back casting method. The same set of stations will be used to estimate a before and after situation. The first time the choice model will be run in a dataset without the new station(s) and the second time the model will be run in a dataset including the new station(s).

Since this procedure is done for 'new' stations opened between 2005 and 2014, the actual impact these stations had on demand abstraction and station choice is known as well. For validation purposes the results from the model can therefore be compared with the actual changes in station choice.

4. MODEL ESTIMATION

4.1 ACCESABILITY INDICATOR

The aim for this variable is to generate a measure for each station that is explanatory for the connectivity and accessibility of that station compared to all other stations in the network. These variables are partly based on the “rail service quality indicator” as estimated in the paper of Debrezion, Pels and Rietveld (2009).

The variable will account for the fact that the attractiveness of a station is not only determined by factors such as the frequency, station quality or direct station environment factors at the station in question. Attractiveness is also determined by how well interconnected the station is in the rest of the network. A good interconnected station means that the generalized journey time (in terms of in-vehicle travel time, waiting time and transfer penalty) to other stations is low. At the same time the number of potential reachable activities (e.g. jobs, shops, and restaurants) should be as large as possible. In other words: a station should give access to as many as possible opportunities for activities while the (generalized) journey time to these activities is as low as possible.

A good example for demonstrating this issue can be found in Apeldoorn. Besides the main station of Apeldoorn there are two other stations in the city, *Apeldoorn de Maten* and *Apeldoorn Osseveld*, which are separated from each other by less than 1000 metres. Socio-economic circumstances in terms of population served are similar as both stations serve about 4000 people within 2500 metres. Also the type of service is the same: both stations are served by all-service trains twice an hour. The difference in the number of passengers per station however is quite large. *De Maten* handles about 600 passengers a year whereas *Osseveld* handles a 1000 (Source: NS, 2010) passengers a year.

It is most likely that the difference in the number of passengers should therefore be contributed to the fact that *De Maten* is situated on a side branch of the main railway line only linking the relative small town of Zutphen with Apeldoorn (blue line). Traveling to other cities requires a transfer. *Osseveld* (red line) on the contrary, is situated on a main railway line linking *Osseveld* with a large number of larger cities with a direct connection. The most important direct connections per station can be seen in Figure 6. The rail accessibility index should therefore take this effect into account.

In this research two definitions of network accessibility will be used as identified by Porta & Schreurer (2006):

- **Closeness centrality:** Is defined as an inverse weighted function of generalized journey time between the station in question and all other stations in the network.
- **Efficiency or Straightness Centrality:** This indicator is defined as the ratio between the travel distances by train and the shortest distances by road transport from the station in question to all other station in the network.

As not every relation from one station to another has the same importance, all station-to-station relations should be weighted accordingly. This weight (or importance) of the relation ij for station i is based on the probability this trip will be made and the size of the destination station j . The probability a trip will be made is derived from the trip distribution of a gravity model. The potential of a station is

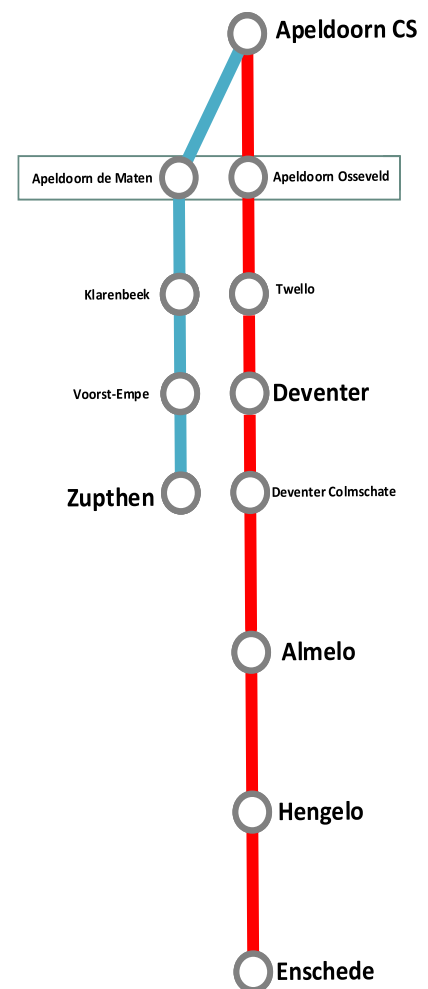


Figure 6: Overview of the most important direct connections of the two sprinter stations in Apeldoorn (right).

expressed as the total number of passengers the destination station j is receiving as observed in the gravity model.

Estimating the Accessibility Index

In order to be able to calculate the accessibility indices, the distribution of trips across the network should be known. A gravity model was therefore estimated using Omnitrans traffic modelling software with the following lognormal form:

$$F_v(Z_{ijv}) = \alpha v * e^{(\beta_v * \ln^2(Z_{ijv} + 1))}$$

With parameters:	α	Mode specific parameter. Only applies in multimodal networks.
	v	Mode (in this case train)
	β_v	Parameter to be estimated based on average travel time
	F_v	Indicating the distribution function for mode v
	Z_{ijv}	The impedance between station i and j for mode v

The parameter β was estimated based on the method from *modelling transport* (Ortuzar and Willumsen, 2009). The initial estimations was based on the given value that an average trip by train takes 38 minutes (Source: Dutch Railways) with access and egress modes excluded. Using $\beta_0 = \frac{1}{C^*}$ a first estimate could be made with C^* as the average travel time as observed by NS.

In the following iterations the formula $\beta_1 = \frac{\beta_0 * C_0}{C^*}$ is used until C^* and C_m have converged enough where C^* is the average travel time as measured by NS and C_m is the average travel time estimated by the model. After 5 iterations β was estimated to be -0.579.

It is also assumed that on short distances, train is less favourable compared to other modes of transport such as bicycles, and other public transport. However when distance increases the train becomes more attractive. Therefore it is expected the majority of the trips will take around 38 minutes, the mean train trip length in the Netherlands. When the trip length grows considerable longer than the average trip length, the probability of making such a trip becomes lower.

After the basic parameters were estimated the gravity function was used to assign trips to the network based on actual passenger counts at all railway stations in the Netherlands as measured between 2010 and 2013. For the assignment phase, a skim matrix was used that represented the actual travel time, waiting time and included penalties for possible transfers. The time period that was modelled consisted of one full working day and therefore the output that was modelled represents the daily flow of passengers on an average working day. The resulting distribution of the number of trips relative to the generalised traveling costs of the trip as calibrated in the model can be seen in figure 5.

Based on the trip distribution data as derived from the gravity model, the weight w_{ij} is defined as the fraction (or probability) of the total number of trips which falls within travel time category c :

$$\delta_c = \frac{\sum_{ij} c_{(GJT(ij))} * T_{ij}}{T_{tot}}$$

With parameters:	δ_c	The weight of the trip between station i and j
	T_{ij}	The number of trips between i and j
	T_{tot}	The total number of trips that were made in the model
	$c_{(GJT(ij))}$	Binary variable which is 1 if $GJT_{(ij)}$ falls in category c and 0 if not.

In total 60 categories c , each with a 5 minute span of travel time, were used. A plot of all 60 categories and the corresponding percentage of the total amount of trips can be seen in Figure 7.

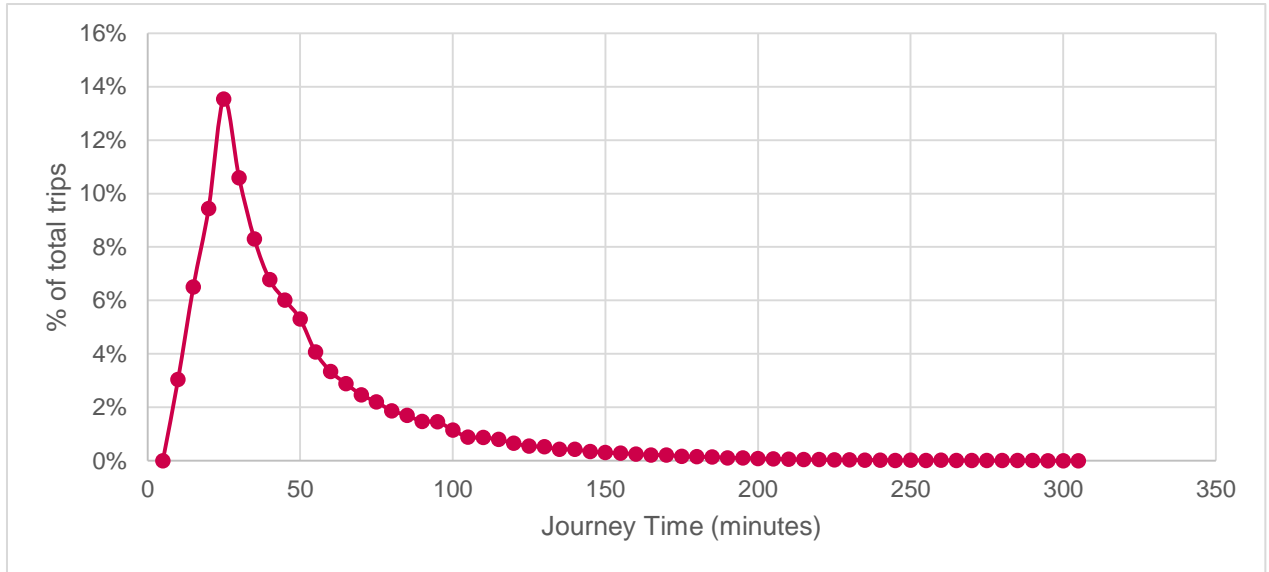


Figure 7: Trip distribution in actual journey time

Next step is to calculate the actual accessibility index using these probabilities. In total three indices will be calculated. The first index will be a basic index B_{ij} based on only the GJT based weight δ_{ij} and the potential of the destination D_j :

$$BI_i = \sum_{ij} (\delta_{cij} * D_j)$$

This is the basic index. But adaptations of this index will be used in order to include certain aspects such as the number of transfers (closeness centrality indicator) or the distance over road compared to the distance by rail (straightness centrality indicator).

The following formula was used to calculate the second indicator (the Closeness Centrality Index) based on the closeness centrality definition:

$$CCI_i = \sum_{ij} (\delta_{cij} * D_j * \frac{1}{c_{ij}+1})$$

With parameters:

CCI_i	The closeness Centrality Index of station i
δ_{cij}	The probability of taking a trip from i to j
D_j	The total number of passengers arriving at station j
c_{ij}	The number of transfers needed to get from i to j

In a similar way the Straightness centrality is also calculated:

$$SCI_i = \sum_{ij} (\delta_{cij} * \frac{L_{road(ij)}}{L_{rail(ij)}} * D_j)$$

With parameters:

SCI_i	The Straightness Centrality Index of station i
$L_{rail(ij)}$	The Distance from station i to j by train
$L_{road(ij)}$	The Distance from station i to j over road
δ_{cij}	The probability of taking a trip from i to j
D_j	The total number of passengers arriving at station j

As a final adaption of the output, all the stations indices were normalized in order to be better able to compare the stations with each other. Utrecht central station was taken to be the reference station as this station is often seen as the best connected and centrally positioned station in the Netherlands and plays a central role in the Dutch railway system.

An overview of the final indices' distribution can be seen in figure 8. These figures show the indices plotted against each other. It can be noticed that in general a higher index score for one index means a higher score for the other. In other words: There is a positive correlation between the two indices. However, in some occasions There can be a large difference between the two indicators. A station scoring high on SCI and low on CCI means that this station is poorly accesable by rail but by car as well.



Figure 8: CC index plotted against the SC index

The top 5 best scoring and the worst scoring stations of both indices can be found in Table 6. As expected the best scoring stations of both indices are all in the Randstad area while the lowest scoring are all outside the Randstad area.

It should be kept in mind that this index is weighted, based on trip data from the gravity model. In the Randstad region there are many relatively well inter connected stations with high frequencies and a large amount of large potential destinations. Therefore the short trips between these stations are weighted relatively high compared to trips to other destinations resulting in a high index score.

SCI	Station	CCI	Station
Highest scoring stations			
1,00	Utrecht Centraal	1,00	Utrecht Centraal
0,97	Amsterdam Bijlmer ArenA	0,99	Schiphol
0,92	Gouda	0,92	Duivendrecht
0,90	Schiphol	0,90	Amsterdam Bijlmer ArenA
0,88	Breukelen	0,90	Leiden Centraal
Lowest scoring stations			
0,03	Veendam	0,03	Workum
0,03	Workum	0,03	Hindeloopen
0,02	Hindeloopen	0,03	Koudum-Molkwerum
0,02	Koudum-Molkwerum	0,02	Stavoren
0,02	Stavoren	0,00	Geerdijk

Table 6: Top 5 of best and worst scores for the CCI and SCI indices

The lowest 5 scoring stations consist of local train stations in the North of the Netherlands are found in Table 6. They all score very low as they are poorly connected with the rest of the national rail network and relative far away (in terms of generalized journey time) from regional hubs as well.

Station	CC Index	SC Index
Stavoren	0,02	0,02
Leeuwarden	0,11	0,12
Zwolle	0,37	0,40
Amersfoort	0,78	0,85
Amsterdam Centraal	0,87	0,76
Utrecht CS	1,00	1,00
Eindhoven	0,42	0,42
Geerdijk	0,07	0,00
Marienberg	0,13	0,19
Hardenberg	0,09	0,14

Table 7: Overview of various intercity and sprinter stations and their corresponding SCI and CCI index scores.

It becomes clear that in general a station scores high in the SCI index when the station in question is located along one or more very (spatially) direct railway corridors between major stations. As the size (number of passengers) of the station itself is not taken into account in the index this station doesn't have to be another major station. It even can be a sprinter station as well. Therefore a sprinter stations such as Breukelen has a high score in this index because of its position right between various large stations in Amsterdam and Utrecht central station. (Semi-)Intercity stations in-between multiple large stations are scoring high as well such as Amsterdam Bijlmer-ArenA (between Amsterdam CS, Utrecht CS) and Gouda (between Rotterdam CS, Utrecht CS and Den Haag CS).

For the CC index the number of transfers required to reach a station becomes more important and therefore Stations with more direct connections will score higher regardless of the distance compared with car. The station of Geerdijk for example is in the CCI the lowest scoring station while more than 50 stations scored worse than Geerdijk in the SCI. This means that although compared with doing the same trip by car, this station offers a reasonable direct connection, but the number of transfers that has to be taken in case this journey is made by train is quite high. On the contrary, Leiden CS does offer direct connection to all major train station in Rotterdam, The Hague, Amsterdam and Utrecht. However the distance too many of them is quite long. Therefore Leiden is able to have a good score in the CCI but a lower score for the SCI.

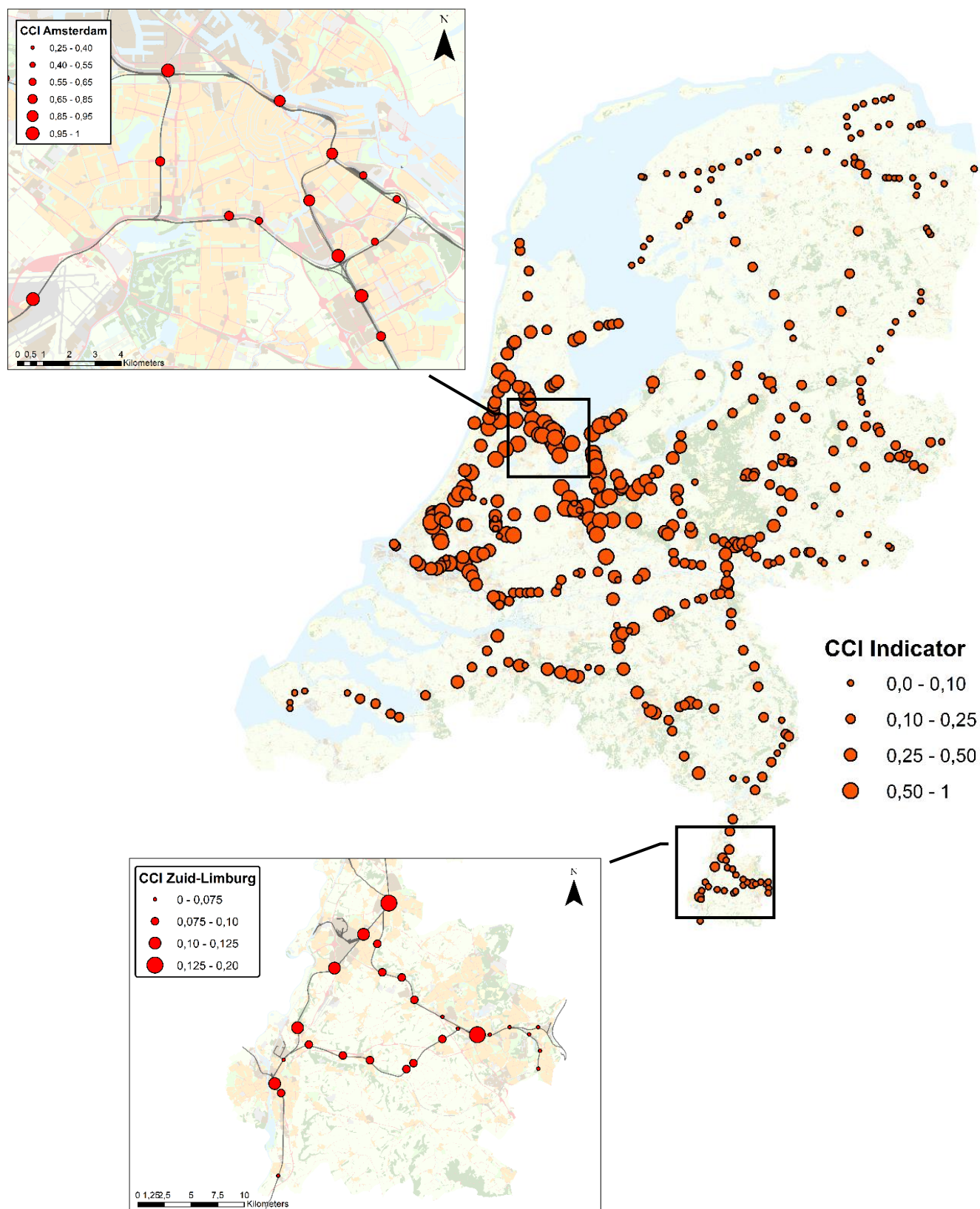
Some other results and comparisons between stations of these indices can be found in Table 7. It becomes clear that major IC stations score much better (above 0.1) compared to local train stations. However, if for example the stations of Geerdijk, Marienberg and Hardenberg are taken out for a closer look it becomes clear that the indicator gets a grasp on minor connectivity differences as well as these three stations are all located close to each other.

Marienberg scores best as at this is the point two lines meet. Therefore this station offers travel opportunities in three directions. The other two stations only offer two travel directions. But Hardenberg is located on a line between two major cities (Emmen and Zwolle) whereas Geerdijk only offers a direct journey to Almelo.

When looking at the national scale (Map 1) it can be noticed that station in the Randstad area (Intercity & Sprinter) are returning better scores than station outside the randstad area. This has to do with the fact that the Randstad area has direct (intercity train) connections with most other major cities in the Netherlands and the fact that most economic activities find place in this part of the country. Therefore even a small sprinter train station along the railway line Amsterdam-Rotterdam has the potential to reach more places within a certain (generalized) journey time compared to a intercity station in the far North or south. Therefore randstad station generally score higher compared to non-Randstad stations.

And finally, when having a look on the scores of the two stations this chapter started with: The local railway line station Apeldoorn de Maten has a CC index score of 0,16. The station on the line of national importance, Apeldoorn Osseveld, scores 0,18. This indicates the index takes the connectivity issues into account.

Map 1: Overview of the CCI indicator on a national scale and in the Amsterdam & South Limburg region



4.2 DISTANCE DECAY FUNCTIONS

The goal of this chapter is to develop a way to define the station catchment area and derive a weighted number of passengers from this catchment area to be used in the regression analysis. This way the number of inhabitants living within the catchment area is taken into account in the regression analysis in a more realistic way resulting in a better prediction of the total number of passengers that will use a station.

In other direct demand ridership models the catchment of a railway station was often defined on the basis of an all-or-nothing approach. A buffer is created around the station of for example 5 kilometres and every inhabitant within this area is considered a potential user of the station question. Everyone outside this buffer zone is not taken into account in any way. This method is very simple and straightforward to use but not realistic and therefore also less reliable.

Intermediate catchment definitions were developed which include a bit more detail. Instead of one buffer zone multiple buffer layers are projected around the station. Depending on in which zone or circle a person lives, a weight (the amount of passengers that can be expected per inhabitant) is applied. People living further away from the station are less heavily weighted than people living close to the station. The model used by the Dutch railways works according to this concept.

Another option to increase the quality of the catchment definition is to use the network distance instead of the crow flight distance for determining the buffers. Especially in cases when spatial barriers are present in the direct station environment such as rivers, highways or the railway itself, using the crow flight distance would result in an overestimation of potential users of a station. Using the network distance instead solves this problem.

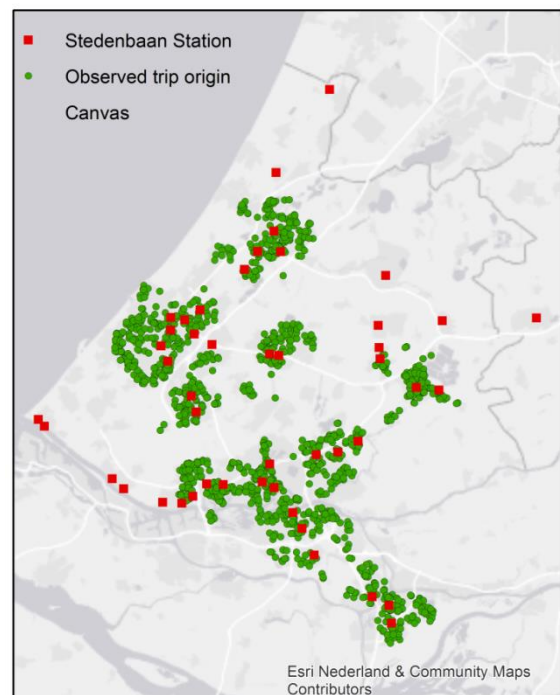
However, even when all of the described model improvements are implemented, all of these catchment definitions are still discontinuous. They all work according to the all-or-nothing principle and make no distinction in the type of station is made although this could influence the catchment area as well. In order to improve the catchment definition a continuous decay function is needed which can estimate the number of potential passengers in an area, depending on the network distance from the station and type of station.

Survey

In order to improve the catchment definition of a station, disaggregated trip data is needed. An online survey conducted in 2013 by the University of Twente in the province of Zuid-Holland was therefore used. This was a survey originally meant to do research on transit oriented development in the South wing Randstad area and was completed by a total of around 1500 respondents. The recruitment was based on three criteria:

- Frequency of traveling by train: a good mix of frequent and non-frequent train users
- Residential location: only people from the South wing Randstad area were selected
- Type of departure station: all six station types should be represented

Out of all 1566 respondents the average age was 54. This average age might be somewhat higher than in reality and therefore the results might be biased. It can be the case that older people tend to travel less far in order



Map 2: overview of all stedenbaan stations and trip points of origin

to get to the station than younger people. The distribution of all respondents over access modes and station types can be seen in Appendix III. The collected data consists of a revealed preference and a stated preference part. For the calculation of the distance decay functions only the revealed preference data was used.

The first step in generating the decay functions is to determine the maximum distance people are willing to travel to a railway station. Since the trip origin and departure station is known for most observed trips (Map 2) the distance between the origin and departure station can be derived. This was done in ArcGIS by geo-referencing every zip code and observed departure station to the map location. Secondly a network was made with use of the network analyst tool in ArcGIS based on the Open Street Map dataset.

The “Nationaal wegen bestand Nederland” was used as well. However, results from this network were unreliable as the measured distances were longer than could reasonably be expected. The problem was the fact that many minor walkways, cycle ways and passages for pedestrians were missing in this dataset. Therefore it was chosen to continue with the open street map dataset which, in general, had these types of roads included as well.

In the process of the network building special attention needs to be given for the attachment points of the object location to the network. The GIS software attaches a point (a station) only to one link in the network. However certain stations can be accessed from both sides of the railway tracks. These connecting tunnels at near or at the stations are often already present in the open street network dataset. However, sometimes they are missing and have to be included manually.

Table 8: Basic statistics of the access distance to the station per station type and mode

Source: Open Street Map		Distance to station (Access)				
		Minimum	Maximum	Standard Deviation	Mean	Count
Station Type	1	155	13750	2741	4709	282
	2	250	11952	1709	2882	381
	3	151	12683	2738	3075	174
	4	177	7300	1815	2775	160
	5	116	16275	2045	2246	250
	6	267	14523	2815	3059	53
	Total	116	16275	2405	3195	1307
Access Mode	Car_Passenger	687	12973	2989	4412	57
	Car_Driver	653	13750	2474	3839	77
	Bus, tram or metro	971	12683	2395	4576	240
	Cyclist	429	9763	1553	2432	230
	pedestrian	116	16275	2243	1714	213
	Total	116	16275	2535	3174	825

Using these inputs, the total distance of all observed routes from origin to departure station could be estimated (Table 8). An important remark is that in this process the routes of all respondents are calculated in the same way although the corresponding access mode might differ. Therefore the routes calculated for car- and public transportation-users might be smaller than they are in reality. Also it could be argued that people are not always taking the shortest path during their trip to the station. They might prefer another route or are simply unaware of the fact that another route might be shorter.

The results in table 2 were also cleaned from outliers. There were some instances of cases which returned corresponding routes of over 20 kilometres by foot for example. As these cases are most likely an error caused either by a wrong understanding of the question in the survey or data processing. Therefore all modes trip distances longer than 20 kilometres were eliminated.

A final remark is that some data was not suitable for estimating the trip length as either the postcode was missing or the departure station was unknown. This data was excluded as well. As can be seen in Table 8, the total number of observations of which the access mode is known is roughly 500 cases less than the total number of observations. This limits the possibility of including access modes in the distance decay curves. Therefore no distinction is made between different types of access modes in the distance decay curves.

Following the same method also the egress distances from the station to the destination were estimated (Table 9). The number of observations however is limited and therefore only one distance decay curve for the egress side will be estimated.

Station Type	minimum	maximum	mean	count
1	172	9887	2360	136
2	249	4967	2288	54
3	144	8849	3210	48
4	7614	8704	8159	4
5	609	2744	1147	16

Table 9: Basic statistics of the egress distance from the station per station type

The results in Table 8 & Table 9 do show that in other research the catchment area of a station reaches quite a bit further than distance thresholds commonly used in literature for train stations. Especially the stations of type 1 seem to have a large catchment area. The distances per mode are like expected: Un-motorized modes have in general a limited range (around 2 kilometres); Motorized modes tend to have a longer range (up to 14 kilometres).

Now the route lengths are known the next step is to determine the probability a passenger choosing for a station comes from a certain distance from that station. In other words: the fraction of observation per distance band to the total number of observations. This was done by counting the number of observation in all distance bands according:

$$P_{bt} = \frac{n_{bt}}{N_t}$$

With: P_b Probability of a passenger of a station coming from distance band b
 n_b The number of observation in distance band b for type t
 N The total number of observations for type t

The distance bands are bands of 500 metres with the first band measuring from 0 to 500 metres and continuing to 14500 to 15000 metres. This was done for all six station types and for all station types combined.

However this probability cannot yet be used as a weight in the regression analysis. Problem is that the survey is only a sample of train users. Nothing is known about the total amount of train users and how many train users the survey is representing of this total amount. It can only be assumed that the survey is representative for the spatial distribution of people choosing to travel by train.

Therefore a final adaption has to be made before decay function can be estimated. The survey was conducted for departure station of which the total number of boarding passengers is known. Therefore the total number of expected passengers per distance band can be calculated according to:

$$E_{tb} = B_t * P_{bt}$$

E_{tb} *The expected number of passengers of station type t originating from band b*

B_t *The total number of Passengers at station of type t*

P_{bt} *The probability of a passenger coming from distance band b*

Now by dividing the number of expected passenger with the actual number of inhabitants in band b the number of passengers per inhabitant is derived:

$$W_{bt} = \frac{E_{tb}}{I_{bt}}$$

W_{bt} *The number of passengers per inhabitant in band b for station type t*

E_{tb} *The expected number of passengers of station type t originating from band b*

I_{bt} *The total number of inhabitants in band b for station type t*

This way, for every station type the amount of passengers per distance band was derived. On the basis of these values distance decay functions were estimated. All scatterplots of these weights and the proposed decay functions are found in Appendix IV.

The decay function for the egress side of the trip was estimated with the same method. However, instead of the number of inhabitants this function was estimated with the number of jobs. This function should therefore be read as the number of train trips can be expected per job.

In order to choose the right function type multiple functions have been fitted with the data. The function type with on average the best fit for sprinter stations was chosen based on R^2 . The decision to only take sprinter train station into account is because it is assumed that for future new train station sprinter train stations are the most common type.

The data was tested with linear, logarithmic, exponential and quadratic function types (Table 10). The logistic function proved to have the best fit for sprinter train stations. In table 11 all estimated decay functions, constants and the corresponding R^2 can be found. All station types have good fits except station type 3 and in a lesser extend type 6. Too few observations for these station types might cause the bad fit. A solution would be to get more observations of users of these stations in order to be able to get a more detailed image of the geographic distribution around these stations types and perhaps make the calibration distance bands smaller for even more detail. Unfortunately this data is not available.

Table 10 & 12: decay function per station type (left) and various tested function types (right).

Type	Cons*ln(x)	R ²
1	-0,134ln(x) + 1,2534	0,75
2	-0,13ln(x) + 1,1973	0,96
3	-0,068ln(x) + 0,6048	0,86
4	-0,055ln(x) + 0,4809	0,90
5	-0,065ln(x) + 0,5402	0,89
6	-0,158ln(x) + 1,3303	0,65
IC	-0,065ln(x) + 0,5525	0,95
Sprinter	-0,129ln(x) + 1,1903	0,91
Egress All	-0,193ln(x) + 1,6753	0,89
Egress IC	-0,401ln(x) + 3,4839	0,89
Egress Spr	-0,116ln(x) + 0,9807	0,93

Function type	Average R ² Sprinter
$W_{bt} = ax + b$	0,55
$W_{bt} = ax^2 + bx + c$	0,83
$W_{bt} = ax^{-b}$	0,85
$W_{bt} = aLN(x) + b$	0,93

The final estimated distance decay curves are shown in Figure 9 & Figure 10 on the next page. In figure 10 also the combined graphs of Sprinter, Intercity and egress can be found. As can be seen, there are significant differences between the station types:

- Station type 1: Has the largest catchment of all stations. As this type of station is also the largest station type usually well connected with public transportation (including metro and tram) this is not surprising. This type of station was already described as a well-connected stations with a (inter)national focus. The corresponding estimated decay function proves that a station of this type therefore also has a city wide and regional catchment area.
- Station type 2: This station type has a large catchment area as well although slightly smaller. This is because this station type is usually to be found in city centres of middle to large sized cities. The decay function proves that the focus of these stations lay mainly in the city centre as the catchment area and the overall trip production for this station is lower than that of type 1 stations.
- Station type 3: Type 3 stations are well connected stations usually on the edge of larger cities. Since these stations serve as satellite stations of type 1 and 2 stations these stations have considerable smaller catchment areas. However the catchment area is larger than those of type 4 and 5 stations because of the good connectivity theses stations have because of limited intercity service and feeding bus, tram and metro lines.
- Station type 4: These are sprinter stations and therefore the initial attractiveness is not even half of that of type 2. However, as these stations are located near the centre of small towns and villages the initial attractiveness is still higher than for example stations of type 5.
- Station type 5: These are small suburban stations near the edges of a city. The catchment area of this station will therefore be limited to the residential area directly around the station. The catchment area is therefore small. As the service level is usually low with only sprinter trains the initial attractiveness is very low.
- Station type 6: This type of station is usually to be found a considerable distance away from the centre of a town in open area. Therefore it would be expected that the initial attractiveness is fairly low but the catchment area is still quite large. However, the initial attractiveness is unexpectedly with 0.7 passengers per inhabitant quite high while the catchment area is similar to type 4 and 5. However this was also the curve with the lowest R² and these results should be handled with care.

Egress: The catchment area for the egress function is with a threshold of less than 6000 metres is considerably smaller than for the access functions of type 1 & 2. This can be expected as the mode car or bike is not that often used as an egress mode. This means un-motorized modes (especially walking) and public transport are the only remaining modes to leave the station.

There also seems to be a large difference between the egress functions of intercity and sprinter stations. This can be mainly contributed to the fact that type 1 and 2 stations usually have a better public transportation connectivity offering passengers a better connection to their final destination.

Figure 9: Graphs of the distance decay function per station type

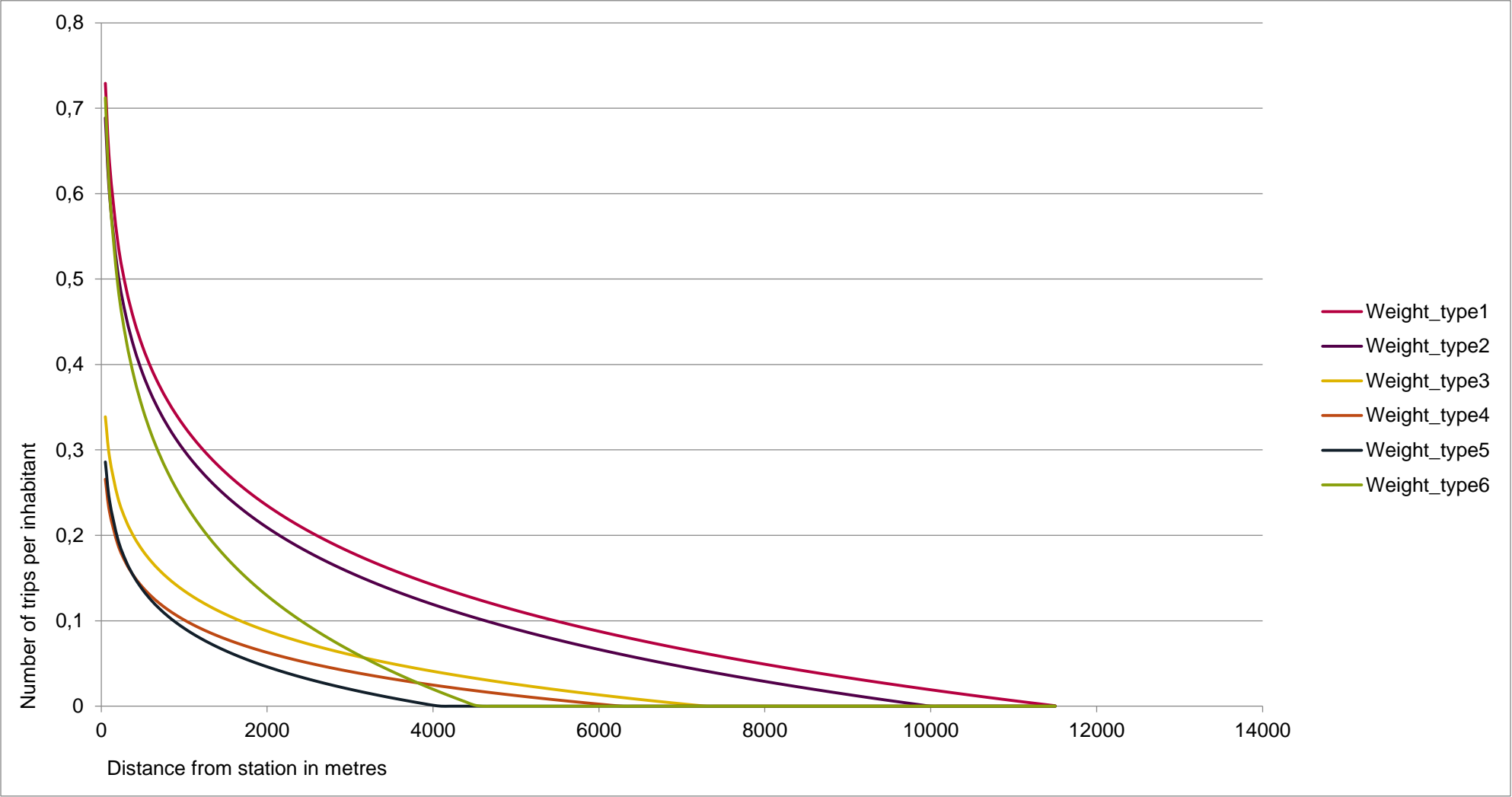
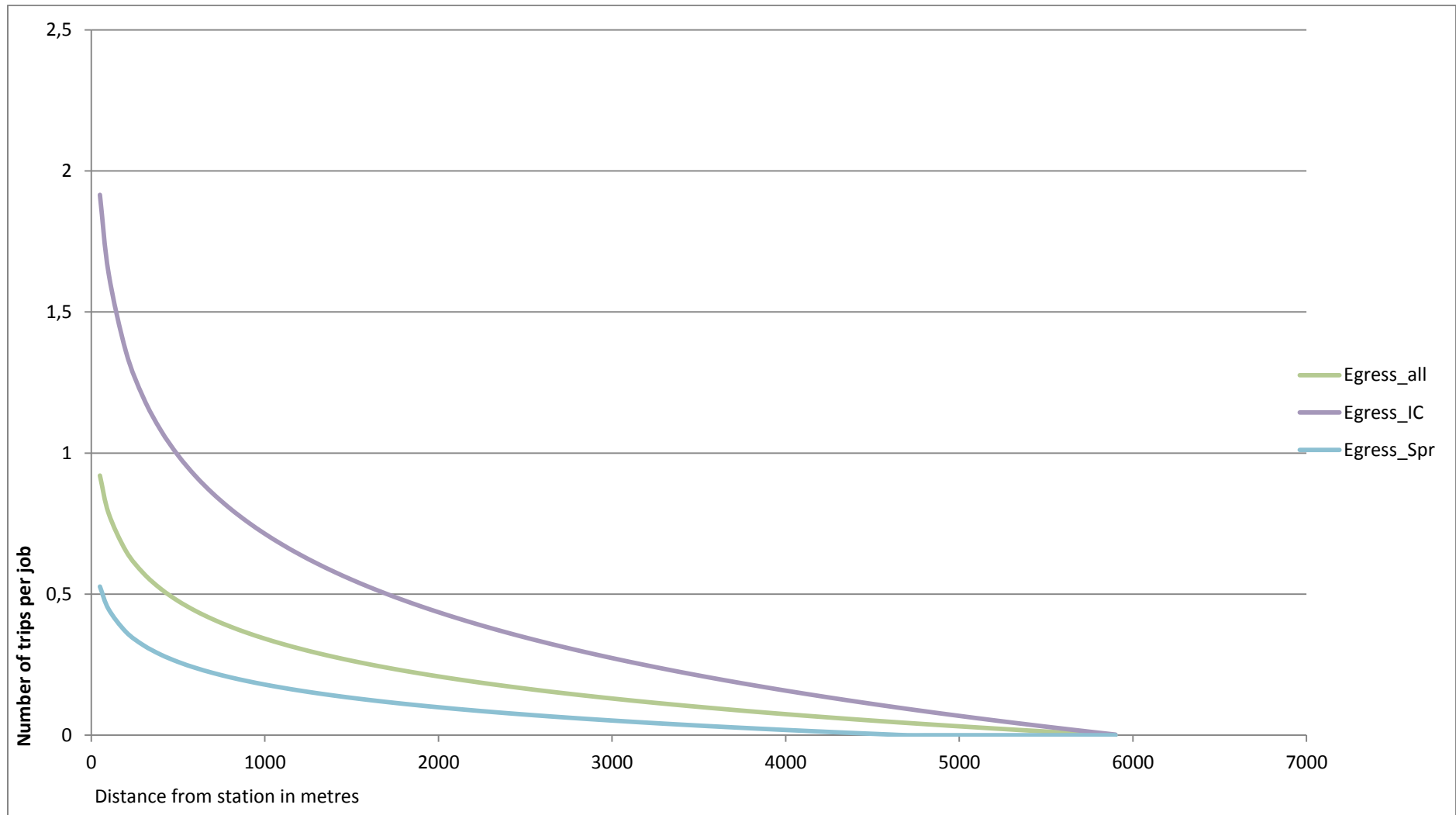


Figure 10: Distance decay functions for sprinter, intercity stations & for egress.



4.3 STATION CHOICE MODEL

The station choice model is meant as a tool to assign passengers to a station on the basis of distance, accessibility and other possible attributes. Especially in situations where multiple stations compete for the same traveller, a station choice model can help to identify the catchment areas of each station separately.

Choice set

The choice set is dependent on what the application of the model will be. In this case the goal is to make an estimation of what station people will choose. As this station choice model assumes passengers already decided to travel by train, the (main) mode choice component is not important.

All choices in the choice set have to be significantly different from each other, there are two ways to base the choice set on:

- A choice set with the x number of closest stations based on distance.
- A choice set with the closest station for each station type (types 1 till 6) or certain types combined (i.e. closest sprinter (type 3, 4, 5 or 6) and intercity station (type 1 & 2)).

However, in both occasions there are potential problems with the choice sets. Within the first choice set a problem could arise when a new station is added to the choice set. In case this new station will become the first ranked station, all other stations will descend one place in rank. Since each rank has a separate utility function with its own beta parameters, this means that the utility of a choice can change while no variables of the station itself have been altered.

This problem was already identified in multinomial logit model estimations using a choice set of ten stations: *"In all cases the ASC for the 10th ranked station is less negative than the ASC for the 9th ranked station, meaning that a change in ranking from 9 to 10 would, all else being equal, mean that the probability of choosing that station would increase. In an attempt to overcome this problem, nested logit models were tested which split the choice dataset into 'local' and 'railhead' stations but these had a far inferior fit, as did aggregate choice intervening opportunity models"* (Blainey & Evans, 2011).

Problem with the second type of choice set is that in case stations of the same type are competing (for example Apeldoorn de Maten and Apeldoorn Osseveld, only one will be represented in the choice set. This is not acceptable when implementing the model and therefore this choice set is considered less suitable for this goal. A rank based choice set based on only the distance is therefore a better solution because all stations within a certain range are represented regardless of the station type.

Despite of the results of the research of Blainey & Evans (2011) the preferred model type is model containing choices based on rank since this type of model will at least include all stations available regardless of two station have the same station type or not. However, in order to make a more conceptually pleasing model, an intermediate model type combining the two choice sets was made as well.

Table 11: Number of observations per rank

Rank	Frequency	Percent
1 st	785	56,3
2 nd	231	16,6
3 rd	95	6,8
4 th	73	5,2
5 th	34	2,4
6 th	22	1,6
7 th	22	1,6
8 th	11	,8
9 th	3	,2
10 th & beyond	119	8,5
Total	1395	100

A total of ten stations as presented in the research of Blainey & Evans (2011) might be too much in this case. Considering the observations from the Stedenbaan survey (Table 11) a maximum of five choices is a better choice since the category “other” , representing people choosing a station ranked 5th or more, is limited to 15% while at the same time keeping the number of cases of the 1st till the 5th rank is at an acceptable level of at least more than 34 cases in each choice group.

Passengers choosing a station beyond a 15 kilometre threshold are considered outliers and are not included in the calibration of this model. In order to limit the amount of computing time, the choice categories 1st till 5th ranked stations are limited to a 15 kilometre threshold as well. This means that in case there are only 2 stations available in a 15 kilometres, the 3rd, 4th and 5th ranked stations are not available for this case.

Variable selection

Multiple variables are tested in the station choice model. Three types of variables have been identified for possible inclusion in the model:

1. Rail accessibility
2. Station accessibility
3. Origin characteristics

The rail accessibility is considered as the accessibility a passenger has when already arrived at the departure station. Variables in this group are the frequency and number of lines served by intercity and/or sprinter trains & the accessibility indices (CCI and SCI).

The second category “station accessibility” is defined by variables explain the quality of getting to the departure station from the point of origin by different modes. Station accessibility by public transport is defined by the number of lines and frequency of bus, tram and metro lines. Accessibility by bike includes the availability of (guarded) bicycle parking facilities. Accessibility by car can be measured in the number of parking spaces and/or the availability of park & ride facilities at a station. For all modes the total distance from origin to departure station is part of the accessibility as well.

One problem with mode specific variables (especially variables related to cycling) is that the relevance of these variables might differ depending on the distance the station is from the point of origin. The availability of guarded bicycle parking is more relevant if the station is 4 kilometres away than it is when the station is 14 kilometres away (Figure 11). However, in both occasions it is possible the station is 2nd ranked. Therefore this effect is not captured within the choice set.

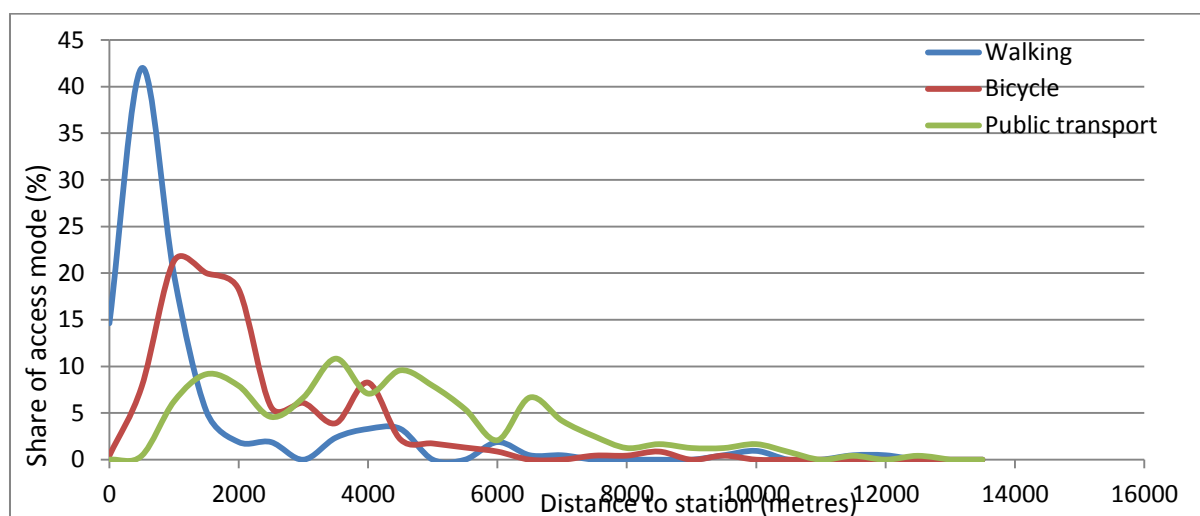


Figure 11: Share of pedestrians, cyclists, and public transport users plotted against the distance to the station. Share of car users and other modes excluded. Source: Stedenbaan survey.

In a station choice model with a rank based choice set this poses a problem, especially for cycling related variables. In order to overcome the problem the variable 'Bike' (availability of guarded bicycle parking) is weighted according to the share of bicycle users at that distance (see figure 8). Therefore the availability of a guarded bicycle shelter should result in a higher utility when the corresponding station is 2 kilometres away compared with the same station when it is 10 kilometres away. Although a similar effect is in place for other modes this is not as big of a problem as there are no mode specific variables in place (pedestrians) or the spread in mode share is considerably large (motorized modes).

The third category consists of variables which are defined by the postcode of origin (the level on which the model will be applied). This can be car ownership, income or the size of certain age groups. Since the category 'other' (stations chosen which are ranked 6th or beyond) is expected to be high when there are almost no or relative many stations to be found within 15 kilometres, two other variables on postcode level were defined: Average distance to all other available stations within 15 kilometres & total number of stations within a 15 kilometre radius.

Final MNL choice models

Based on the different choice sets and based on the conceptual value of the model, three different MNL station choice models have been estimated (see Table 12):

Model	Final Log_likelihood	R ²
MNL Model 1 (rank based)	-1082.42	0.504
MNL Model 2 (type & rank based with flexible parameters)	-1119.90	0.416
MNL Model 3 (type & rank based with fixed parameters)	-1158.96	0.399

Table 12: Overview of MNL station choice model results

MNL choice model 1

This first model is the model with the best overall fit. However, because of the large amount of variables included it is also the most complicated model. The model is based on station ranks and it includes, if available, five stations (within 15 kilometres) and the category 'other'. All parameters of the model can be seen in Appendix 3.

The Following utility functions were used for MNL station choice model 1:

$$V_{rank1} = ASC_1 * one + Frequency1 * Freq1 + Distance1 * Dist_1 + Dist1 * Ratio_1 + BTM1 * NOL_BTM1 + RAIL_acces1 * IND_1 + Bike1 * Bike_Park1$$

$$V_{rank2} = ASC_2 * one + Frequency2 * Freq2 + Distance2 * Dist_1 + Dist2 * Ratio_2 + BTM2 * NOL_BTM2 + RAIL_acces2 * IND_2 + Bike2 * Bike_Park2$$

$$V_{rank3} = ASC_3 * one + Frequency3 * Freq3 + Distance3 * Dist_1 + Dist3 * Ratio_3 + BTM3 * NOL_BTM3 + RAIL_acces3 * IND_3 + Bike3 * Bike_Park3$$

$$V_{rank4} = ASC_4 * one + Frequency4 * Freq4 + Distance4 * Dist_1 + Dist4 * Ratio_4 + BTM4 * NOL_BTM4 + RAIL_acces4 * IND_4 + Bike4 * Bike_Park4$$

$$V_{rank5} = ASC_5 * one + Frequency5 * Freq5 + Distance5 * Dist_1 + Dist5 * Ratio_5 + BTM5 * NOL_BTM5 + RAIL_acces5 * IND_5 + Bike5 * Bike_Park5$$

$$V_{rank_other} = ASC_other * one + Frequency6 * Freq6 + Distance6 * Dist_1 + Dist6 * Ratio_6 + BTM6 * NOL_BTM6 + RAIL_acces6 * IND_6 + Bike6 * Bike_Park6 + other_St * other_St_1$$

Of the variables described earlier a total of 6 have been included in the model plus an additional variable which is only included in the utility function for 'other stations'. Since the model is fully flexible a different coefficient is estimated for every utility function. Therefore the value of this coefficient is describing the importance of this variable for the corresponding choice.

Distance: The variable distance is the most important variable in this model for especially the second ranked model. For the 3rd, 4th and 5th ranked stations this variable is slightly less important. Also the distance ratio (distance of station in question as a fraction of distance to closest station) is a significant variable for all ranks meaning that when the relative difference between the distance of the closest and for example a second closest station is small, the utility of this second closest station will increase. In other words: The likelihood that a lower ranked station is chosen depends on the difference in distances of the closest and the lower ranked station. The smaller this distance difference is, the higher the chance a lower ranked station will be chosen

Secondly the frequency coefficient becomes larger for utility functions for lower ranked stations. This means that in order for a lower ranked station to be eligible to be chosen it must have a relative high frequency compared to the closest station. A similar effect is in place for the accessibility index (CCI).

As for bicycle parking facilities, only the closest station is able to increase its utility if one is available. If bicycle storage is available for a lower ranked station this has no effect on the utility of that station.

The number of bus, tram and metro (BTM) lines connecting the station is about equally important for all station ranks. Since this is a motorized mode it makes sense the effect of distance (or ranking) on this variable is limited.

The choice option 'other' is a somewhat different choice category. This is mainly because there are only two situations in which it is expected that the category other will attract a large share of the passengers:

1. In situation where there are **many** stations available combined with relative **short** distances
2. In situation where there no or **few** stations available combined with relative **long** distances

The first situation will only occur in dense urban areas (near Amsterdam, Rotterdam and The Hague) when there are more than 5 station available within a relative short distance from the point of origin. When none of these five stations is distinctive in any way, the model will assign a large share of passengers to the category 'other'.

The second situation occurs in rural areas. In case there is no station available within 15 kilometres, 100% is assigned to the category 'other'. In case there are stations available but only in a long distance the share of other will somewhat decline though it will remain relatively high. However, when there is one or more stations available at a reasonable distance the 'other' category will become small.

In the model this mechanism is simulated by taking the average of all variables of all choice options (i.e. the average frequency, average CCI index, and average distance). In a situation where no choice option is significantly better than any of the other choices, the utility of the category 'other' will become relative high.

Despite of the relative good fit of this model, the problem is that after inclusion of a new station the utility (and thus final demand) might actually become higher. In reality it is however highly unlikely that ridership of a station will become higher when a new station is being opened.

MNL Choice Model 2

Because of the conceptual constraint of MNL choice model 1 a new model was estimated. This model is in contrary to the previous one based on both station ranks as well as station type. The choice set consists of the two closest intercity stations, the two closest sprinter stations and the category 'other'. It is expected that this model will suffer less from the conceptual problem of a station getting a higher utility when its rank becomes lower due to the inclusion of a new station.

The Following utility functions were used for MNL station choice model 1:

$$V_{\text{rank_1IC}} = \text{ASC_6} * \text{one} + \text{Distance1} * \text{Dist_IC} + \text{Frequency_IC} * \text{Freq6} + \text{BTM_IC} * \text{NOL_BTM6} + \text{Bike_IC} * \text{Bike_Park6} + \text{Index_IC} * \text{CCI_6}$$

$$V_{\text{rank_2IC}} = \text{ASC_7} * \text{one} + \text{Distance2} * \text{Dist_IC2} + \text{Frequency_IC} * \text{Freq7} + \text{BTM_IC} * \text{NOL_BTM7} + \text{Bike_IC} * \text{Bike_Park7} + \text{Index_IC} * \text{CCI_7}$$

$$V_{\text{rank_1sprint}} = \text{ASC_8} * \text{one} + \text{Distance3} * \text{Dist_Spr} + \text{Frequency_Spr} * \text{Freq8} + \text{BTM_Spr} * \text{NOL_BTM8} + \text{Bike_Spr} * \text{Bike_Park8} + \text{Index_Spr} * \text{CCI_8}$$

$$V_{\text{rank_2sprint}} = \text{ASC_9} * \text{one} + \text{Distance4} * \text{Dist_Spr2} + \text{Frequency_Spr} * \text{Freq9} + \text{BTM_Spr} * \text{NOL_BTM9} + \text{Bike_Spr} * \text{Bike_Park9} + \text{Index_Spr} * \text{CCI_9}$$

$$V_{\text{rank_Other}} = \text{ASC_other} * \text{one} + \text{Distance5} * \text{Dist_other} + \text{other} * \text{other_St}$$

The parameters of this model were estimated as followed:

Parameter	Beta parameter	T_score
Rho-Square	0.416	
Final Log liklihood	-1119,9	
ASC_6	0.00	
ASC_7	-1.04	-3.88
ASC_8	-0.633	-1.33
ASC_9	-2.38	-3.53
ASC_other	-0.935	-1.38
BTM_IC	0.0169	2.75
BTM_Spr	0.0373	5.94
Bike_IC	0.00	
Bike_Spr	0.117	4.79
Distance_IC_1 st	-0.000824	-15.47
Distance_IC_2 nd	-0.000502	-9.78
Distance3_Spr_1st	-0.000982	-14.17
Distance4_Spr_2nd	-0.000769	-6.28
Distance_other	-0.000632	-7.43
Frequency_IC	0.0440	6.52
Frequency_Spr	0.00	
Index_IC	0.00	
Index_Spr	1.94	2.54
Other	0.128	2.74

Table 13: Overview of the model parameters of Station Choice model 2.

Strong point of this choice set is the fact that now parameters can be estimated separately for sprinter and intercity stations. Therefore, although the same variables are used as in model 1, the coefficients are somewhat different.

The number of bus, tram and metro lines (BTM) is much more important for sprinter stations than it is for intercity stations. A similar effect is in place for the variable describing bike parking facilities (Bike_IC & Bike_Spr) and for the general rail accessibility. This is most likely because these variables are important in describing the difference between intercity stations and sprinter stations, but not for

the difference between two intercity stations since intercity stations always have a high BTM connectivity and bicycle parking facilities available.

The utility function for the category other is simplified a bit but still has the same effect. The category 'other' will receive a relative large fraction in case no other station is standing out in a positive or negative way. Otherwise the 'other' choice option is only marginal.

MNL Choice Model 3

This model was estimated in order to eliminate the conceptual problem of changing utilities due to changing ranks. It is just as model 2 based on a combination of station ranks and types. However, in this model a stations' utility can only be changed when the variables of the station in question are changed. This also results in the fact that demand for existing stations in the choice set can only decrease when a new station is added.

The following utility functions were used for MNL Choice model 3:

$$V_{\text{rank_1IC}} = \text{ASC_6} * \text{one} + \text{Distance_IC} * \text{Dist_IC} + \text{Frequency_IC} * \text{Freq6} + \text{BTM_IC} * \text{NOL_BTM6} + \text{Bike_IC} * \text{Bike_Park_6} + \text{Index_IC} * \text{CCI}$$

$$V_{\text{rank_2IC}} = \text{ASC_7} * \text{one} + \text{Distance_IC} * \text{Dist_IC2} + \text{Frequency_IC} * \text{Freq7} + \text{BTM_IC} * \text{NOL_BTM7} + \text{Bike_IC} * \text{Bike_Park_7} + \text{Index_IC} * \text{CCI}$$

$$V_{\text{rank_1sprint}} = \text{ASC_6} * \text{one} + \text{Distance_Spr} * \text{Dist_Spr} + \text{Frequency_Spr} * \text{Freq8} + \text{BTM_Spr} * \text{NOL_BTM8} + \text{Bike_Spr} * \text{Bike_Park_8} + \text{Index_Spr} * \text{CCI}$$

$$V_{\text{rank_2sprint}} = \text{ASC_7} * \text{one} + \text{Distance_Spr} * \text{Dist_Spr2} + \text{Frequency_Spr} * \text{Freq9} + \text{BTM_Spr} * \text{NOL_BTM9} + \text{Bike_Spr} * \text{Bike_Park_9} + \text{Index_Spr} * \text{CCI}$$

$$V_{\text{rank_Other}} = \text{ASC_other} * \text{one} + \text{Distance5} * \text{Dist_other} + \text{other} * \text{other_St}$$

The following model parameters were estimated for station choice model 3:

Parameter	Beta parameter	T_score
Rho-Square	0,399	
Final Log_liklihood	-1158,96	
ASC_6	0.243	2.39
ASC_7	0.00	
BTM_IC	0.00	
BTM_Spr	0.0255	5.25
Bike_IC	0.00	
Bike_Spr	0.116	5.58
Distance_IC	-0.000659	-15.56
Distance_Spr	-0.00102	-15.55
Distance_other	-0.000669	-8.28
Frequency_IC	0.0623	15.55
Frequency_Spr	0	
Index_IC	0	
Index_Spr	1.53	4.28
other	0.114	3.89

Table 14: Overview of the model parameters of station choice model 3.

Since many of the model parameters are fixed, the overall fit and R^2 of this model is relative low. However, conceptual this model meets the requirements. Separate parameters are estimated for sprinter and intercity stations. However, for the ranks no separate parameters have been estimated.

Model application

In order to demonstrate the quality and applicability of the models, all three models are applied in a situation in which a new station is opened. For this goal the opening of a new station (Leeuwarden Werpsterhoek) in the city of Leeuwarden is use. In Figure 12 the result of each model is found. It shows the change in the fraction of the people per postcode area choosing for the main station of Leeuwarden. It is expected that the opening of Leeuwarden Werpsterhoek will cause a decrease of demand for the main station of Leeuwarden, especially on the south side of the city. In other areas no change is expected.

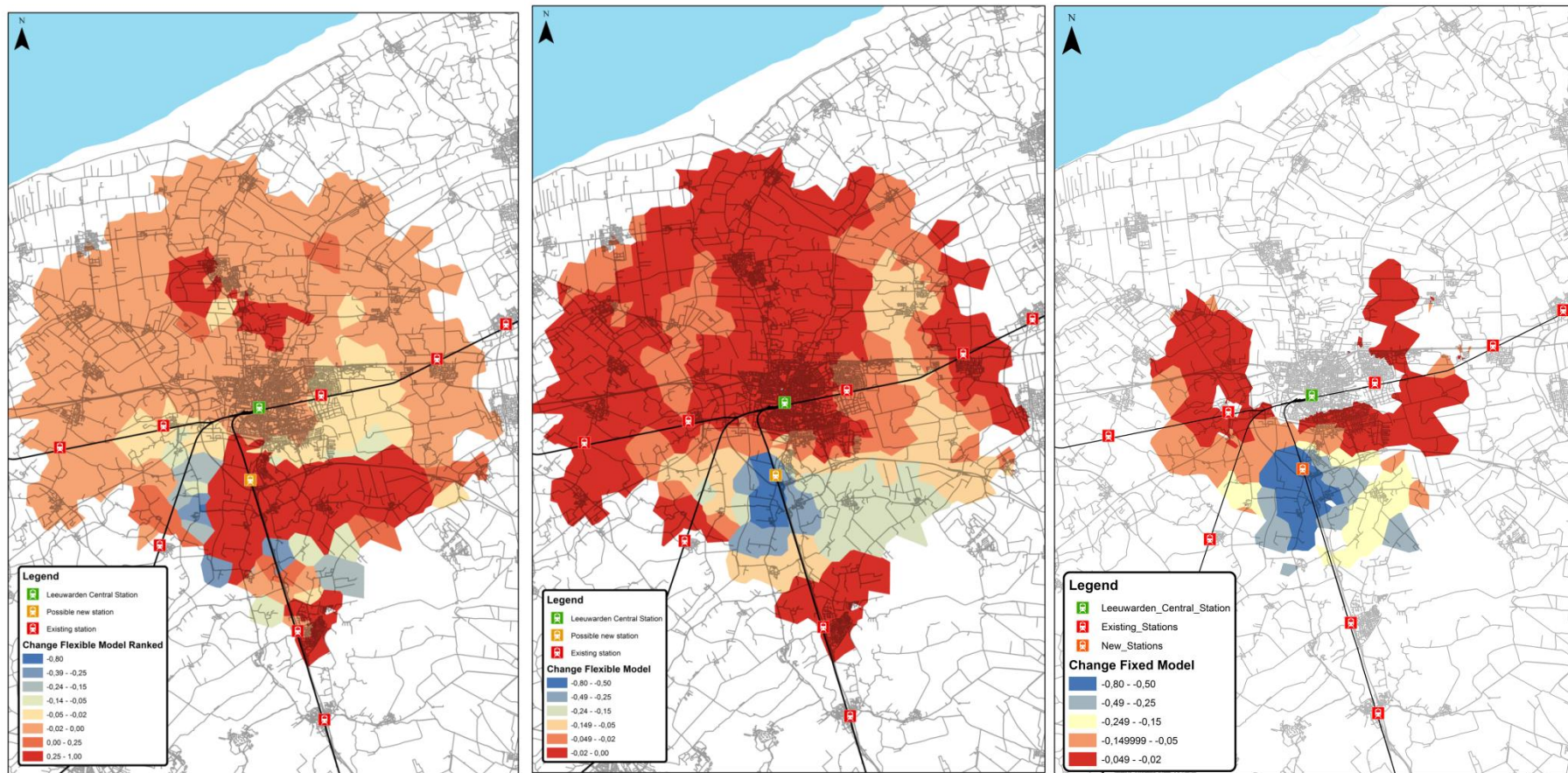


Figure 12: Change in demand after the opening of Leeuwarden Werpsterhoek as modelled with (from left to right) choice model 1, model 2 and, model 3.

MNL station choice model 1 returns the worst result in terms of what can be expected despite of being the model with the best fit. Although some areas show a decrease in demand as expected, others show an increase of demand which is not realistic in a situation in which only a new station is added and none have been closed. Also the increase and decrease of demand is not consistent as relative large changes in demand can be observed north of the city as well. This is mainly caused by the problem as discussed before that changing ranks cause a change in utility and ultimately a change in demand of that station. Therefore it can be concluded that this model is not suitable for application in practice.

The other two models, MNL choice model 2 & 3, are performing considerably better. A relatively large decrease of demand is visible in the direct proximity of the new station. Other areas remain relatively untouched. The small decrease of demand in general is contributed to the fact that the overall station density within this area has gone up with one station. This will slightly increase the utility of the category 'other'. However this fraction is marginal (0.02 or less) and will therefore have no or little influence on the final results.

Model 3 however performs the best in terms of what would be the expected result. This model has the smallest affected area due to the opening of the new station. Areas North, East and West of the city are not as affected or not affected at all. At the same time the impact the new station has on the decrease in demand for the intercity station is higher.

Because of these results Model 3 has the preference to be applied in practice since this model returns the intuitive better results. Even though this model has the worst fit of all three models, the results are better when the model is applied.

4.4 INITIAL STATION POTENTIAL

On the basis of the distance decay curves (

4.2 Distance Decay Functions) and the station choice model (4.3 Station Choice Model), new density variables can be calculated. These variables are expected to be better able to explain ridership since these initial potential variables are corrected for the distance decay effect and overlapping catchment areas.

Whereas in literature the total population, number of jobs and, total student enrolment is often used as direct input for the regression analysis, these variables are now used as the starting point for the new density variables.

First of all, the distance from a six digit postcode to its nearest 6 stations is calculated. After application of the corresponding distance decay curves, the total potential for rail transport in this postcode area is estimated. In order to prevent double counting of demand, the station with the highest potential for a postcode area is used as the total potential for this area. This is done for population, the number of jobs and the number of students. After a rail potential for each postcode area is known the station choice model is applied to distribute the total potential over the available stations.

This ultimately results in six new enhanced density variables: The total station potential (Tot_Pot), total potential from the population (Pot_Inw), total potential from the number of jobs (Pot_Job), and the potential for the three education levels (Pot_MO, Pot_MBO and, Pot_HBO). An example of how this works out for stations in the Arnhem-Nijmegen city region is found in Map 3. A full list of every sprinter station and its corresponding potentials are found in Appendix 4.

It becomes clear that there are only a few stations where student enrolment and/or jobs are more important than the population (Arnhem Presikshaaf, Nijmegen Heyendaal). The potential of the majority of the sprinter stations is explained on basis of the population, often by over 75%. This is not

only visible in map 3 but is a national trend as well. Only 38 sprinter stations out of the 388 sprinter stations nationwide have a ridership potential that is explained with at least 50% by student enrolment and jobs.

However, it should be noted that this total potential (or Pot_Tot) is only explaining ridership on the basis of the total population, number of jobs and student enrolment. It is not yet corrected for the general rail accessibility in relation to the rest of the network, frequency, social characteristics etc. This will be done in the regression section.

However, the total potential of a station should be able to make a first impression on the actual ridership that can be expected.

As shown in Figure 13 a linear relationship is visible. However, on an individual station level the over and underestimation can be large. Stations with less than 200 daily passengers are overestimated in many cases. At the same time stations with more than 2000 passengers are underestimated in many cases.

Most likely cause is the fact that variables such as bus feeders, network accessibility and frequency are not yet taken into account. With the application of a regression model the severity of this problem should be reduced.

Map 3: Potential of sprinter stations in the Arnhem-Nijmegen city region.

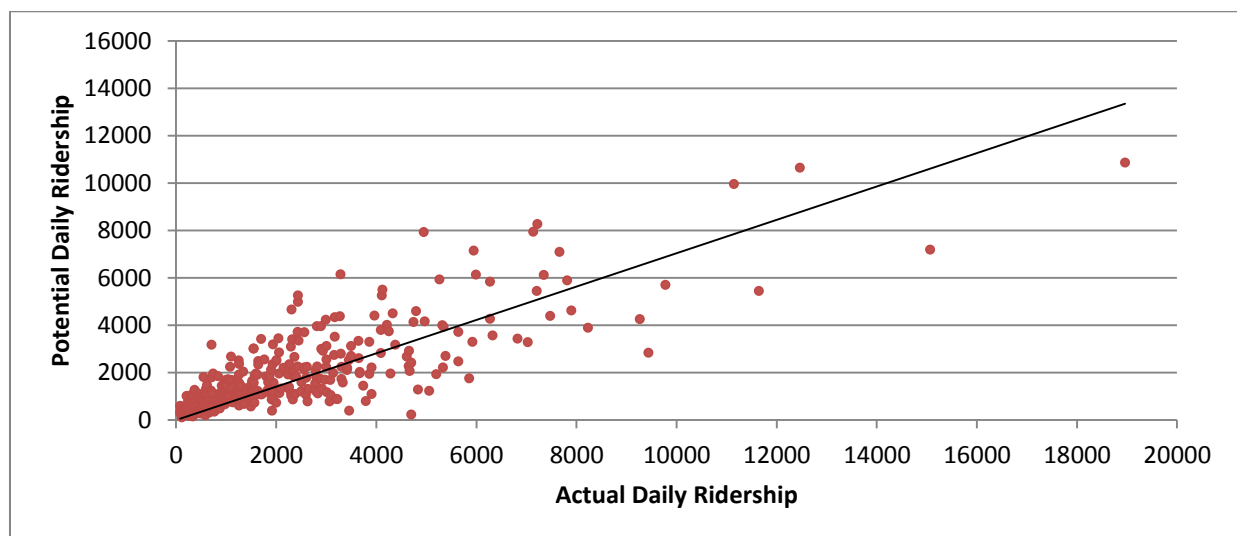
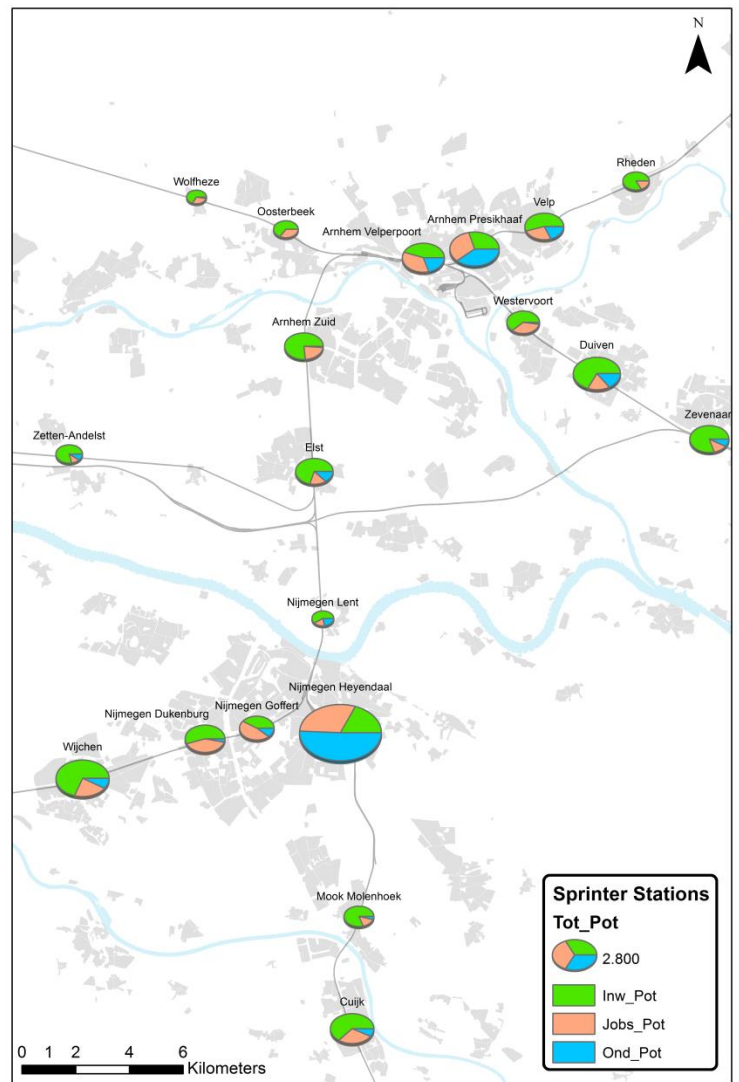


Figure 13: Actual versus Potential ridership demand measured in daily boarding per station

4.5 CORROLATIONS

The regression analysis is the final part in this research. In this part the distance decay curves and accessibility indicators will come together into a model than can predict the ridership of new railway stations. Also new variables will be used in order to improve the regression models.

The basic idea of a multiple regression analysis is to explain the relationship between a variable Y and the variables X1, X2 up to Xn. The final model can be described as:

$$Y = A * B_1X1 * B_2X2 * B_3X3 * B_nXn$$

With parameters:	Y	Dependent variable
	A	The constant or intercept of the model
	B _n	The constant for variable X _n
	X _n	The variable X _n

In section “3.4 Data” an overview of all data sources and corresponding variables was given. Also new connectivity or rail accessibility variables are calculated in section “4.1 Accesability indicator”. Section 4.4 described the calculation of a new population, job and, student enrolment variables. The final step will therefore be to estimate a regression model.

The dependent variable is the daily number of passengers boarding and exiting a train at a specific station in 2013 (Daily_2013). Furthermore, all variables used in the regression need to be ratio, scale or binary variables. Nominal variables should not be used in the regression. A complete list of all variables used in the regression can be found in appendix 2.

The first step in the pre-selection process is to make a selection of variables that are most likely to have a good explanatory value in a final regression model. Secondly the correlation amongst these variables should not be too large.

Variable pre-selection

Before the actual regression models will be estimated, a pre-selection will be made on the basis of the correlation coefficient with the dependent variable. This correlation coefficient (Pearson correlation) is a figure between -1 and 1 where a negative score means a negative correlation.

In total 88 variables are included in the regression. It is expected that some of these variables are not or only limited contributing in explaining ridership at railway stations. These variables will most likely have a small correlation with the dependent variable (daily_2013). Using these variables as input for the regression models is not useful. Others will be highly correlated with the dependent variable and are therefore useful in the regression model. In Table 15 a full overview of all variables and corresponding correlation is found.

Table 15: Correlation of independent variables with the dependent variable (Daily_2013).

Variable	Correlation	Variable	Correlation	Variable	Correlation	Variable	Correlation
AUTO_HH	-0,42	Tour_Bed_Rel	0,18	AV5_BIOS	0,35	P_N_W_AL	0,49
MP_HH_ZK	-0,39	P_Cult	0,20	P_woon	0,35	Overdekt_perron	0,51
P_KOOPWON	-0,36	Detail_horeca	0,21	AV20WARENH	0,35	Streekbus_Freq	0,53
AF_ONDVMB	-0,34	park	0,22	BEDR_AUTO	0,36	PR_Cat	0,54
AF_ONDVRT	-0,34	Design_modern	0,22	AV10WARENH	0,36	BTM_NOL	0,55
AF_BIOS	-0,33	woon	0,22	AV5_PODIUM	0,37	Parking_spaces	0,57
Proximity	-0,33	LUM	0,23	Archit	0,37	Bicycle_parking	0,62
AF_ONDHV	-0,32	Pot_HBO	0,23	Tour_Bus_Abs	0,37	Freq_Tot	0,65
AF_PODIUM	-0,32	Ratio_Students	0,23	AV1_SUPERM	0,37	CCI_2013	0,66
AF_WARENH	-0,29	P_Detailhandel	0,23	SCI_2013	0,37	Freq_BTMT	0,67
P75OUD	-0,26	AV1_CAFE	0,24	AV1_RESTAU	0,39	Tourism_Abs	0,69
P_6574	-0,25	P_NIETACT	0,25	Som_Leisure	0,39	Pot_Jobs	0,73
AF_ATTRAC	-0,24	P_bedrijfs	0,25	Stadsverv_NOL	0,40	Pot_Inw	0,78
Basic_station	-0,23	IC_Partial	0,25	Pot_Onderwijs	0,40	Pot_JobInw	0,80
AF_OVERST_orig	-0,23	IC_NOL	0,26	Freq_Gem	0,40	Tot_Pot	0,81
AF_POP	-0,20	AV10ATTRAC	0,28	Stadsbus_NOL	0,41		
AF_OVERST_new	-0,18	AV5_HOTEL	0,28	HH_GRT	0,41		
M_HH_MK	-0,15	Tram_NOL	0,28	Stadsbus_Freq	0,41		
P_ink_li	-0,10	AV5_WARENH	0,29	AV3_ONDVRT	0,42		
P_LEEGSW	-0,08	P_3464	0,29	Pot_MBO	0,42		
Terminal	-0,05	IC_Freq	0,30	AV3_ONDVMB	0,42		
WOZ	-0,04	Design	0,30	AV1_DAGLMD	0,42		
P_0014	-0,03	Pot_MO	0,31	Som_Shop	0,42		
AF_OPRITH	-0,01	Tram_Freq	0,32	A_PART_HH	0,43		
P_WONV2000	0,03	AV10ONDVMB	0,33	AUTO_TOT	0,43		
Wegverkeersterrein	0,04	AV5_ONDHV	0,33	A_BED_Fin	0,43		
Bijz_NOL	0,05	AV5_ONDVMB	0,33	A_BED_Hor_Handel	0,43		
P_1534	0,07	Metro_Freq	0,33	A_BED_Zak	0,43		
Parking	0,07	Som_Horeca	0,33	Streekbus_NOL	0,43		
sport	0,08	AV5_WARENH	0,33	A_BEDV	0,44		
Ratio_Jobs	0,09	Metro_NOL	0,33	Opp_bebouwd	0,44		
Tourism_Rel	0,11	AV3_ONDHV	0,33	P_HOOGBW	0,45		
GEM_ink_pi	0,12	Bijz_Freq	0,34	Stadsvervoer_Freq	0,45		
cultuur	0,12	AV10ONDVRT	0,34	Sprinter_Freq	0,47		
P_ink_hi	0,13	AV5_ONDVRT	0,34	Bev_DH	0,47		
Ratio_Destination	0,14	Sprinter_NOL	0,34	AUTO_LAND	0,48		
bedrijf	0,14	AV5_ZIEK_I	0,34	OAD	0,48		
Hist_station	0,16	P1P_HH	0,35	Delay_2013	0,48		
Other_St_2013	0,18	AV10_ONDHV	0,35	Bicycle_rental	0,48		

Built Environment

Density variables

The density variables generally have the largest (positive) coefficients of all variable types. The enhanced population, job and student enrolment variables are correlated with a large coefficient. A bit surprising however is the fact that the HBO/university level is the lowest correlated education level after MO (high school) and MBO. Other density variables are all positively correlation such as the number of business (in certain sectors (A_BED_#) and the availability of certain services (AV_#).

LUM (Diversity)

Already in the literature review is was discussed that the variable "LUM" is more explanatory for a more constant flow of passengers during the day than it is for actual ridership. Therefore the positive correlation of only 0.23 could be expected.

Design

All design variables are moderately correlated with a positive sign. Only exception is the binary variable "Overdekt_perron" indicating whether or not the platform is roofed. This variable is highly correlated with ridership. However, it should be noted that large stations with a high ridership level usually offer a more extensive service. This means the facilities are also on a higher level and the chance the platform is covered is much higher. The problem with this variable is thus the fact that it is unclear if it is the cause or the effect of high ridership. The availability of station facilities such as bicycle and car parking facilities are highly correlated as well. However the same problem as with the "covered platform variable" applies here as well.

Socio-economic variables

Car related variables

Strongest negative correlating variable is the average number of cars per household (AUTO_HH). Variables related to car ownership were already identified in literature as a potential ridership explain variable with a negative sign. However, the other car related variables are less correlated but, more importantly, also have positive signs. This is not entirely surprising as the other car related variables are expressed in totals such as the total number of cars (in commercial use). Therefore these variables are more suitable as a measure for population density than they are for car ownership.

Distance to services

Another category of variables which is also behaving as expected are the variables describing the distance to certain services. Especially the distance to high school education (AF_VRT and AF_VMB) has relative high correlation coefficients. But also all other "distance to service" variables have relative high coefficients.

Only exception is the distance to the nearest highway on-ramp. It would be expected that when a highway on-ramp is nearby, the accessibility by car is higher and thus a better alternative at the expense of rail transport. However the correlation coefficient is close to zero. This can be explained by the fact that areas where the distance to a highway entrance is small are often urban areas. In these urban areas, this increase of car accessibility might be counterbalanced by the decrease in car accessibility due to congestion and a higher rail patronage in general.

Age

There is no strong correlation found between age groups and rail ridership. Although a high percentage of people aged above 65 years seems to have a slight negative correlation, the coefficient is only -0.26. Children and young adults have almost no correlation and adults between 35 and 65 have a slight positive correlation. The difference between age groups can be explained by the fact that elderly generally make less trips a day than people who are still working/enrolled in education.

Household composition

Households consisting of multiple people but without children (MP_HH_ZK) are strongly and negatively correlated with ridership. An explanation could be the fact that these social groups often have a higher income compared to other social groups. This higher income results in a higher percentage of car owners. However, since the income variables have far smaller coefficients this explanation might not be the whole picture.

Other household variables are only slightly negatively correlated (MP_HH_MK) or are positively correlated (P1P_HH). It was expected that 1 person households are positively correlated since this group includes students. For households with children it is expected that they would travel by train less often as this group is more likely to own a car.

Income

The income variables are all correlated as expected: low income negatively, higher income positively. However, the fact that the coefficient size is 0.13 at maximum does come as a surprise. A larger coefficient was expected on the basis of literature. But since the basis of the assumption that there is a correlation between income and rail patronage was found in research based in Australia, it is likely that this variable might have worked out less significant in a Dutch context where it might be more common for all income groups to travel by train.

Related to income, the value of houses in the station area (WOZ) is basically uncorrelated with ridership. This is also surprising as based on literature, higher income neighbourhoods often have a decreased demand for public transport. Again, this assumption is based on literature from Australia and the US. In the Dutch case the link between income or house value is therefore less evident.

Network variables

Terminal

The variable Terminal, indicating the end of the line, proves to be uncorrelated with ridership. The theory behind this variable is that it is most likely positively correlated as the last station of the line would have a larger catchment area due to the absence of rail infrastructure beyond this point. In the Dutch practice the network density is much higher and end of the line station are often near the seaside. This reduces the potential hinterland of this terminal station significantly.

Station Accessibility

Furthermore all variables describing the number of lines (NOL) or frequency of either metro, tram, bus or train lines are all positively correlated. Special attention goes out to the total frequency of trains (Freq_Tot) and public transport (BTM_NOL) which have the highest correlation coefficients of all network variables. Also the CCI as estimated in section 4.1 highly correlates with the total ridership of a station as expected. The SCI indicator however only has a correlation of 0.37.

Inter-correlation

Second step is to control for inter-correlated variables. Again, this inter-correlation is checked with the use of the Pearson correlation coefficient. In case the coefficient is too large (with negative or positive sign) only one of the two variables in question can be included in a regression model at the same time. The other variables should be excluded from the model.

Since there are no strict rules in determining when a Pearson score is too high or too low in order to be excluded from the regression. This is depending on the quality of the measurements and the dataset. For this research in general a score higher than 0.7 (or lower than -0.7) is regarded as too correlated and therefore one of the two variables in question should be removed from the regression.

In "Appendix 5: Inter-Correlation between variables" an overview of all variables and the inter-correlation is found. This overview only contains the variables having a correlation coefficient larger than 0.35 regardless of sign. In this section only the most important correlations are discussed.

Built environment

Where the density variables were the variables with the highest correlation with the dependent variable, these are also the variables that are inter-correlated the most. Tot_Pot and the separate variables Pot_Inw and Pot_Job are highly correlated. This means a regression model must contain either the total potential based on residents, jobs and student enrolment (Pot_Tot) or only one of the sub-variables "Pot_Job" or "Pot_Inw". The education variables are not highly correlating with each other and can thus be used at the same time.

Highly correlated with all potential variables is the total bus/tram/metro variable (Freq_BTMT). This is no surprise as a higher potential usually comes with denser local public transportation network as these

areas tend to be more urban. Bicycle storage is also correlated just below the 0.7 threshold. This has similar reasons as guarded bicycle parking facilities are more common in urban areas.

Other density variables including the availability of certain services (AV_#) and the number of businesses (A_BEDV_#) are strongly correlated with each other but not with the “potential” variables. Also the variables OAD and Bev_DH are strongly correlated.

Since the potential variables have the largest correlation coefficient with the dependent variable these variables will have priority when estimating a model. The other density variables should only be included when they still are able to add some explanatory value. However, because of the strong correlation between all density variables chances are that the other density variables are only telling the same “story”.

The design variables are not correlating with any other variable. Only exceptions are the variables PR_Cat and Parking_spaces. This is no surprise since the second variable was calculated from the first one.

Socio-economic variables

The number of non-western immigrants is correlated just below the 0.7 threshold with most other socio-economic variables. That makes this variable less favourable to be used in the regression. The number of cars per household (AUTO_HH) is strongly correlated with P_Koop and P1P_HH.

Network variables

Amongst the network variables, variables related to the same mode are correlated above the 0.7 threshold. For example, NOL_Sprinter and Freq_Sprinter are correlated and so are NOL_stadsbus and Freq_Stadsbus. In general that means that the frequency and the “number of lines version” of a variable cannot be used at the same in case they are both related to the same mode of transport. Variables related to different modes can be used at the same time. If the summarised version of the variable is used underlying variables should not be used.

The CCI indicator is not correlating with any other variables above the 0.7 threshold. However, it does correlate with the SCI indicator just below the threshold (0.69).

Conclusion

This section has provided a pre-selection of variables of which it is likely they will add explanatory value to the dependent variable “Daily_2013”. However, this is not a solid selection but only an indication of which variables should be included for the best results.

As for the variables that will most likely perform well in the regression are the potential variables (Tot_Pot, Pot_Inw & Pot_Jobs), the network variables (CCI, Freq_Tot, BTM_NOL etc.) and some station design variables (Parking_spaces, Bicycle_parking).

Correlation in-between independent variables are discussed as well. The result is that certain variables should not be used in the same regression model at the same time. Especially sub-variables that share a summarized variable (Freq_Tot, BTM_NOL) and mode related variables should not be used at the same time. Special attention also goes out to the density variables which are all more or less correlated. Not all correlated variables have been discussed. Therefore also in the next section the correlation remains an important factor to deal with when estimating the regression models.

4.6 REGRESSION MODELS

The regression was conducted using SPSS statistics software. The first step in the modelling process was to generate two models. The models are calibrated with the use of all type 3-6 stations of which data was available. This includes all sprinter stations in the Netherlands according to the timetable in 2013. This is, excluding the type 1 & 2 stations and the stations used for validation, a total number of 325 stations (or cases) used in the multiple regression analyses.

In order to include a variable into the model the variable has to comply with three criteria:

1. A variable should not be too much correlated with other variables already included in the model. A threshold of 0.7 of the Pearson correlation coefficient was used as a guideline. If a variable is included in the model causing a Pearson correlation higher than 0.7, one of the two variables involved should be excluded. The variable to be excluded will be the variable with the lowest contribution to the model in terms of added r^2 .
2. After the variable complies with criteria 1 the variable should be significant as well. If the significance level has a value above the 0.05 threshold, the variable will not be included in the model.
3. The variable added to the model should make a logical contribution to the model. For example, the variable population should not be included in the model if the coefficient is negative. It cannot be easily understood why a higher total population causes a lower potential ridership. The cause of this problem might be multiple minor correlations or unexpected side effects.

Basic Models

Basic General Model

The first model (Table 16: 1a) estimated is a general model valid for the whole country and for suitable for all sprinter station types. It is aimed to be as simple as possible by only containing variables which are most critical in order to be able to make a good estimation. This model should therefore be easy to apply.

Only five variables have been included in this model including Total potential, the CCI indicator, frequency for sprinter and IC trains and finally the number of bus, tram and metro lines passing the station.

Although the model fit is already high with an R-square of 0.837, this model is not very accurate for especially the smaller stations. Also negative predictions are quite common for these smaller stations. Aim for the next model is therefore to increase the accuracy for smaller stations and decreases the number of negative predications.

Extensive General Model

The extensive model (Table 16: 1b) is estimated in order to improve the predication quality. By including more variables the ridership at stations can be better explained. The increase in variables was partly achieved by including sub-variables instead of the summarized variables. Instead of BTM_NOL now separate "number of line" variables were included. This will increase the flexibility of the model as for example a distinction can be made between regional and city busses.

Instead of five now twelve variables were included in the model. R-square increased from 0.837 to .876. However, since the initial R-square was already relative high, the additional increase by including 7 additional variables is limited.

Also the number of negative forecasts has not decreased significantly although the standard error is reduced by roughly 110 trips. It might be useful to further divide the cases into two new groups. By estimating new regression models separately for each group might improve the estimation quality of the models.

Scatterplots of the basic models that show the relationship between the estimated and the actual ridership are found in Figure 15.

Regional Models

The 306 cases are grouped into two categories: Main line stations and regional stations (See Map 4). Main line sprinter stations are in most cases stations along lines which are part of the main railway network and served by NS.

Regional lines are not part of the main railway network and thus separately tendered. Therefore the regional railway lines are often, but not always, served by other transport companies such as Arriva or Veolia.

Main reason to make this separation is because the regional lines are in most cases railway lines in rural areas. The average speed on these lines is lower than on the main lines and the overall rail accessibility along these lines is lower.

Basic Regional Model

The basic regional model (Table 16: 2a) is, similar to the basic general model, aimed to be as simple as possible and therefore easy to implement.

With only 3 variables an R-square of 0.716 is reached. Just as in the general basic model, the total potential and frequency are the two most important explanatory variables.

The third variable is 'Proximity' describing the relative distance to various (urban) services.

With a positive coefficient it means that a larger distance to the nearest urban centre gives a higher rail ridership. The fact that this variable is significant and contributing to the model fit while the accessibility variable CCI was excluded from the model indicates that for regional railway lines the relative remoteness to the nearest (large) town is more important than the relative rail accessibility to all other stations in the network.

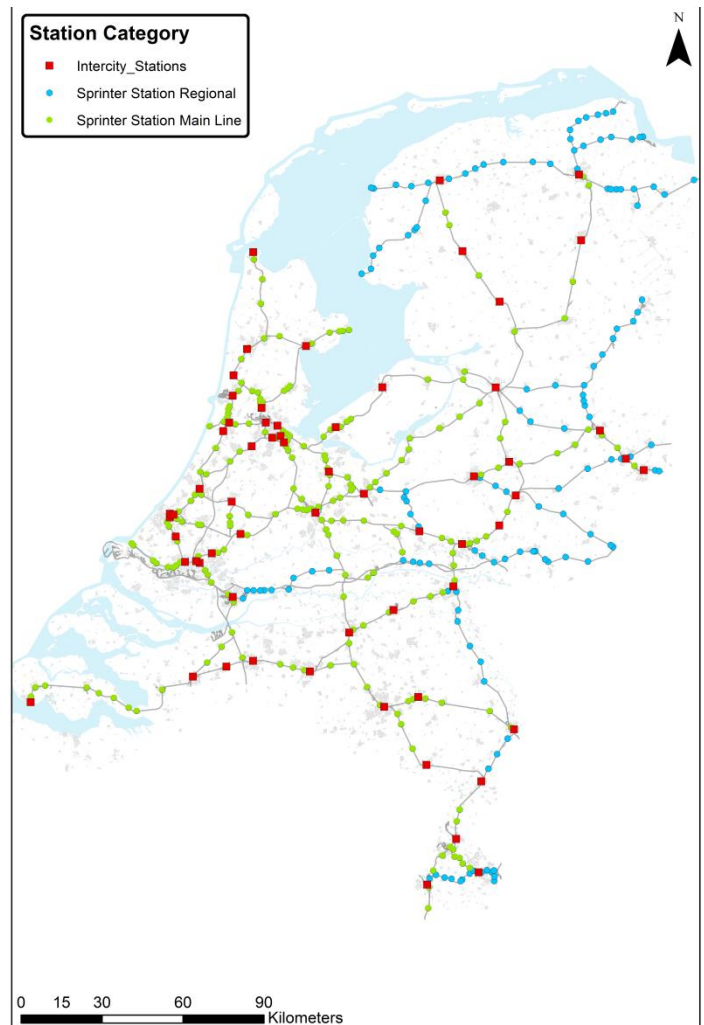
Extensive Regional Model

In an attempt to further improve the model fit also an extensive regional model is made (Table 16: 2b). The R-square was improved from 0.716 to 0.774. This was done by adding four additional variables, the total potential was replaced by three separate variables (Pot_Inw, Pot_MBO and Pot_HBO). Furthermore the CCI indicator and BTM_NOL was included in the model.

Unfortunately the inclusion of additional variables has not lead to an improvement of the estimation for especially the smaller stations. Negative forecasts are more common while at the same time the model fit did increase.

Scatterplots of both the basic as well as the extensive regional model are found in Figure 16.

Map 4: Stations divided in regional stations, main line stations and intercity stations.



Main Line Models

The main line models are regression models based on the stations that are located along the main railway lines in the Netherlands and are all served by NS. For similar reasons as the estimation of the regional models, estimating separate main line models might result in more accurate estimations.

Basic Main Line Model

The basic main line regression model (Table 16: 3a) consists of four variables: Total potential, CCI indicator, total train frequency and, the number of bus/tram/metro lines. Adjusted R-square is 0,798.

The coefficients of total frequency (Freq_Tot) and number of lines (BTM_NOL) are much larger in the main line model indicating a dependence on connectivity. Also the rail accessibility indicator has a larger coefficient compared to the regional models.

Extensive Main Line Model

The extensive version of the main line models (Table 16: 3b) consists of eight variables. The additional variables are all sub variables of summarized variables that were present in the basic main line model. R-square was increased from 0.798 to 0.817. Considering the fact that four additional variables were included this is not a large increase in model fit.

However, the flexibility in this model was increased. Where in the basic model only one coefficient is available for all feeding/competing modes, in the extensive model a separate coefficient was estimated for each mode. The differences in the size of the coefficients are large. The number of tram lines has a large positive contribution to ridership according to this model while city buses are having a negative effect.

Scatterplots of the main line models are found in Figure 14.

Table 16: Overview of all regression model results

1a. Basic General Model						2a. Basic Regional Model						3a. Basic Main Network Model								
Variables	Unstandardized Coefficients		Standardized Coefficients		Sig.	Variables	Unstandardized Coefficients		Standardized Coefficients		Sig.	Variables	Unstandardized Coefficients		Standardized Coefficients		Sig.			
	B	Std. Error	Beta	T			B	Std. Error	Beta	T			B	Std. Error	Beta	T				
(Constant)	-	1419,80	129,73		-10,94	0,000	(Constant)	-	734,53	204,40		-3,59	0,000	(Constant)	-	1531,48	207,90		-7,37	0,000
Sprinter_Freq	231,25	22,95	0,29		10,08	0,000	Tot_Potentie	0,71	0,05	0,79		15,61	0,000	CCI_2013	2174,03	504,47	0,16		4,31	0,000
IC_Freq	408,28	35,51	0,31		11,50	0,000	Freq_Tot	116,92	25,85	0,22		4,52	0,000	Tot_Potentie	0,65	0,06	0,47		10,83	0,000
CCI_2013	3263,72	354,72	0,25		9,20	0,000	Proximity	96,55	35,07	0,13		2,75	0,007	BTM_NOL	57,53	11,53	0,21		4,99	0,000
Tot_Potentie	0,62	0,05	0,45		13,45	0,000							Freq_Tot	293,53	29,93	0,38		9,81	0,000	
BTM_NOL	38,81	8,35	0,15		4,65	0,000														
Model Summary						Model Summary						Model Summary								
N	Model Sig.	R	R Square	Adjusted R Square	Std. Error of the Estimate	N	Model Sig.	R	R Square	Adjusted R Square	Std. Error of the Estimate	N	Model Sig.	R	R Square	Adjusted R Square	Std. Error of the Estimate			
307	0,000	0,916	0,840	0,837	1005,361	119	0,000	0,858	0,735	0,728	556,02	191	0,000	0,896	0,802	0,798	1193,409			
1b. Extensive General Model						2b. Extensive Regional Model						3b. Extensive Main Network Model								
Variables	Unstandardized Coefficients		Standardized Coefficients		Sig.	Variables	Unstandardized Coefficients		Standardized Coefficients		Sig.	Variables	Unstandardized Coefficients		Standardized Coefficients		Sig.			
	B	Std. Error	Beta	T			B	Std. Error	Beta	T			B	Std. Error	Beta	T				
(Constant)	-	1262,97	133,13		-9,49	0,000	(Constant)	-	944,63	195,58		-4,83	0,000	(Constant)	-	1635,50	205,82		-7,95	0,000
Sprinter_Freq	201,56	21,11	0,25		9,55	0,000	Freq_Tot	108,76	23,65	0,21		4,60	0,000	Pot_Joblnw	0,67	0,07	0,40		9,65	0,000
IC_Freq	232,65	38,18	0,18		6,09	0,000	Pot_Inw	1,14	0,10	0,63		11,77	0,000	CCI_2013	3041,02	522,29	0,22		5,82	0,000
CCI_2013	2942,79	345,00	0,23		8,53	0,000	Pot_MBO	1,79	0,45	0,19		4,00	0,000	Sprinter_Freq	257,00	30,33	0,35		8,47	0,000
Pot_Inw	0,84	0,09	0,31		9,71	0,000	Pot_HBO	0,42	0,17	0,11		2,44	0,016	IC_Freq	374,25	43,66	0,32		8,57	0,000
Pot_MBO	1,41	0,41	0,08		3,43	0,001	CCI_2013	1878,58	710,78	0,12		2,64	0,009	Tram_NOL	271,54	83,89	0,12		3,24	0,001
Pot_HBO	0,75	0,22	0,07		3,48	0,001	BTM_NOL	14,99	7,49	0,11		2,00	0,048	Stadsbus_NOL	-160,64	56,94	-0,11		-2,82	0,005
Streekbus_NOL	33,93	7,49	0,13		4,53	0,000	Proximity	77,07	32,36	0,11		2,38	0,019	Streekbus_NOL	58,64	11,14	0,21		5,27	0,000
Stadsverv_NOL	351,73	64,08	0,14		5,49	0,000							Pot_Onderwijs	1,12	0,31	0,12		3,64	0,000	
Stadsbus_NOL	-194,22	46,19	-0,11		-4,21	0,000														
IC_service	953,52	290,81	0,09		3,28	0,001														
Bicycle_parking	700,76	216,48	0,10		3,24	0,001														
Parking_spaces	3,08	0,62	0,13		4,96	0,000														
Model Summary						Model Summary						Model Summary								
N	Model Sig.	R	R Square	Adjusted R Square	Std. Error of the Estimate	N	Model Sig.	R	R Square	Adjusted R Square	Std. Error of the Estimate	N	Model Sig.	R	R Square	Adjusted R Square	Std. Error of the Estimate			
307	0,000	0,936	0,876	0,871	894,279	119	0,000	0,895	0,802	0,789	489,175	191	0,000	0,909	0,827	0,819	1140,623			

Figure 15: Basic (left) and extensive (right) general model scatterplots (estimated vs actual ridership)

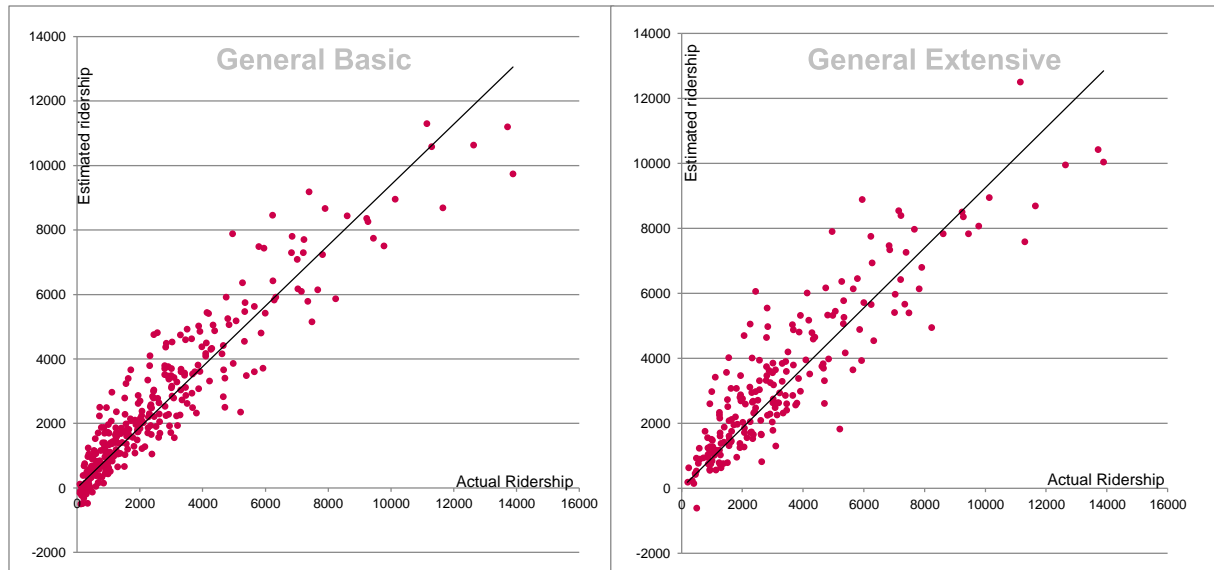


Figure 16: Basic (left) and extensive (right) regional Models scatterplots (estimated vs actual ridership)

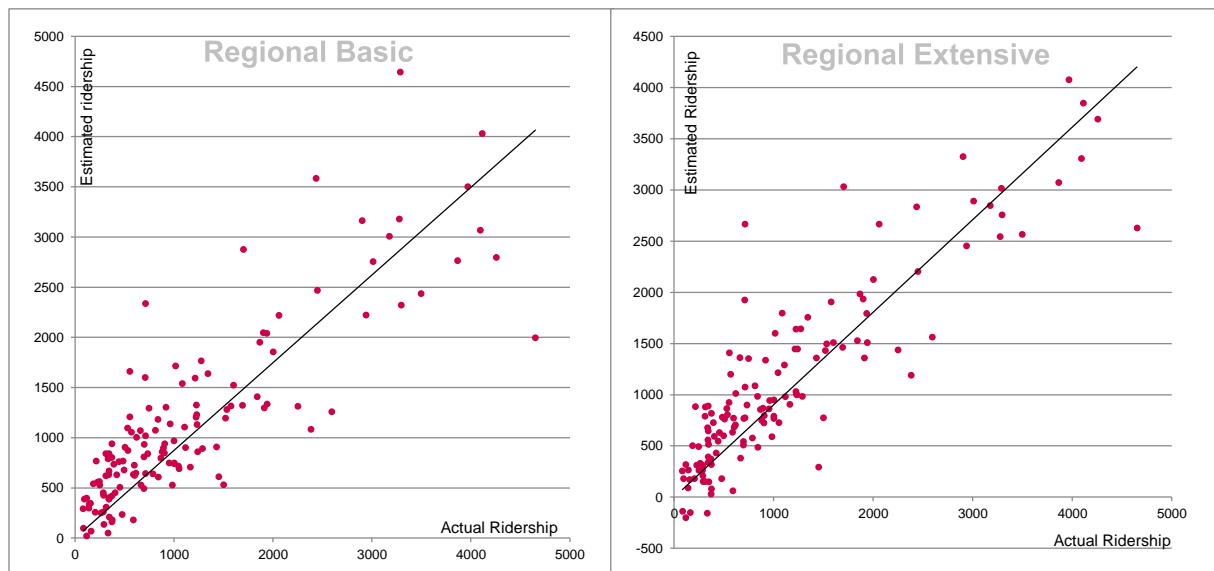
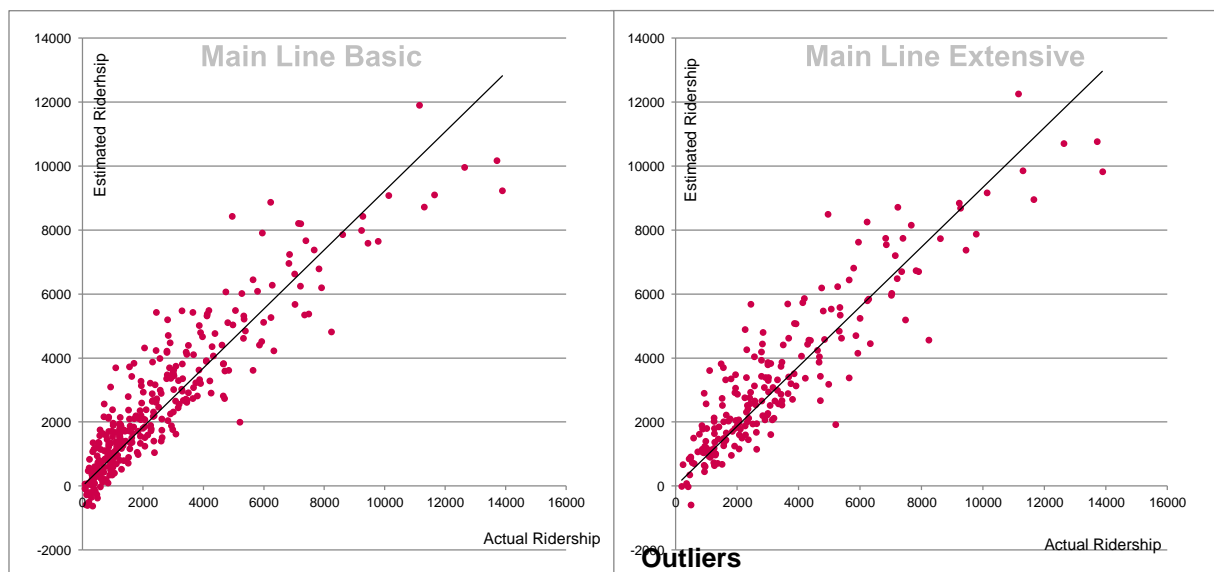


Figure 14: Basic (left) and extensive (right) Main Line model scatterplots (estimated vs actual ridership).



As can be noticed in Figure 14, 15 and 16, basically all of the models which have been estimated contain outliers. That means that ridership for these stations cannot be estimated on the basis of the variables which are included in the current models. The variables that could improve the ridership estimation for these stations might not be available/measured or are too specific and thus only apply for this single station.

Therefore the overall model fit might improve significantly when these outliers are removed from the regression. When removing cases from the regression it also means that the model is no longer valid for all cases. However, the ridership estimation for the remaining cases might be improved significantly.

In order to detect outliers, an analysis has been done on the residues (predicted minus actual ridership count). Cases which have a prediction error of more than three times the standard error are excluded from the next regression step (Table 17).

Table 17: Stations with at least 3 times the standard error per model.

General Basic	General Ext.	Regional Basic	Region Ext.	Main Basic	Main Ext.
Zandvoort	Zandvoort	Zevenaar	Zevenaar	Den Haag I.v. NOI	Culemborg
Culemborg	Woerden		Leerdam	Veenendaal Centrum	
Alkmaar Noord	Alkmaar Noord				
Den Haag I.v. NOI					

Most stations from Table 17 are stations where unmeasured variables have a relative large influence. Ridership at station Zandvoort for example, was underestimated by all regression models. Reason for this underestimation might be the fact that this station, which is located near a popular beach, is attracting a significant amount of passengers with touristic motives all year round. These passengers are not taken into account in the model since their influence on general ridership levels is generally low and thus all touristic related variables returned insignificant. However, in this specific case it has a significant effect on total ridership. Thus it is better to leave this case out of the equation

After the exclusion of the outliers the regression models were rerun with the same variables. The standard error and R-square of every model was increased significantly (see Table 18).

However, the removal of the outliers also brought some changes to the models. First of all the variable IC_service in the extensive general model was no longer significant at could therefore be removed from the model. This makes sense since two semi-intercity stations (Alkmaar Noord and Woerden) were removed from the analysis. Also the CCI indicator and the number of bus, tram and, metro lines were no longer valid variables for the extensive regional model and were thus removed from the regression.

In “Appendix 6: Correlation of Final regression models (minus Outliers)” all correlations in the models are found.

Conclusion

Six different regression models are estimated in this section. Also, all six of them are improved by removing outliers. Overall, the fit of these models is high (0,716 or higher). The general extensive model scores best with an R-square of 0.88. However, these models are not yet calibrated for potential geographic sensibility for certain variables. Also these models still need to be validated and tested before anything can be said which model is best applicable in practice.

Table 18: Overview of all regression models after removal of outliers.

1a. Basic General Model (minus outliers)						2a. Basic Regional Model (minus outliers)						3a. Basic Main Network Model (minus outliers)					
Variables	Unstandardized Coefficients		Standardized Coefficients		Sig.	Variables	Unstandardized Coefficients		Standardized Coefficients		Sig.	Variables	Unstandardized Coefficients		Standardized Coefficients		Sig.
	B	Std. Error	Beta	T			B	Std. Error	Beta	T			B	Std. Error	Beta	T	
(Constant)	1141,88	118,85			0,000	(Constant)	-774,72	206,79			0,000	(Constant)	-1202,01	194,72			0,000
Tot_Potentie	0,64	0,04	0,51	15,56	0,000	Tot_Potentie	0,72	0,04	0,85	16,78	0,000	Tot_Potentie	0,65	0,05	0,52	11,91	0,000
IC_Freq	325,76	34,38	0,26	9,47	0,000	Freq_Tot	102,95	24,55	0,20	4,19	0,000	BTM_NOL	50,15	10,34	0,21	4,85	0,000
Sprinter_Freq	177,90	21,59	0,25	8,24	0,000	Proximity	117,18	34,72	0,17	3,38	0,001	CCI_2013	2867,25	486,52	0,23	5,89	0,000
BTM_NOL	32,07	7,33	0,13	4,38	0,000							Freq_Tot	219,29	30,30	0,29	7,24	0,000
CCI_2013	3384,91	320,34	0,30	10,57	0,000												
Model Summary						Model Summary						Model Summary					
N	Model Sig.	R	R Square	Adjusted R Square	Std. Error	N	Model Sig.	R	R Square	Adjusted R Square	Std. Error	N	Model Sig.	R	R Square	Adjusted R Square	Std. Error
303	0.000	0,92	0,85	0,85	859,78	118	0.000	0,877	0,77	0,76	497,34	189	0.000	0,9	0,81	0,81	1040,7
1b. Extensive General Model (minus outliers)						2b. Extensive Regional Model (minus outliers)						3b. Extensive Main Network Model (minus outliers)					
Variables	Unstandardized Coefficients		Standardized Coefficients		Sig.	Variables	Unstandardized Coefficients		Standardized Coefficients		Sig.	Variables	Unstandardized Coefficients		Standardized Coefficients		Sig.
	B	Std. Error	Beta	T			B	Std. Error	Beta	T			B	Std. Error	Beta	T	
(Constant)	1034,83	121,40			0,000	(Constant)	-973,93	179,77			0,000	(Constant)	-1406,28	195,13			0,000
IC_Freq	235,91	32,08	0,19	7,35	0,000	Freq_Tot	105,26	20,93	0,21	5,03	0,000	CCI_2013	3713,11	505,09	0,30	7,35	0,000
Sprinter_Freq	164,69	19,81	0,23	8,31	0,000	Proximity	110,59	29,31	0,16	3,77	0,000	Pot_Onderwijs	0,99	0,28	0,12	3,60	0,000
CCI_2013	2864,36	314,04	0,25	9,12	0,000	Pot_Inw	1,35	0,08	0,76	17,96	0,000	Pot_JobInw	0,63	0,06	0,42	9,91	0,000
Streekbus_NOL	27,69	6,67	0,11	4,15	0,000	Pot_MBO	2,11	0,36	0,24	5,89	0,000	Streekbus_NOL	55,81	10,12	0,22	5,52	0,000
Stadsverv_NOL	411,71	55,04	0,19	7,48	0,000	Pot_HBO	0,54	0,14	0,15	3,75	0,000	Stadsbus_NOL	-158,19	51,29	-0,12	-3,08	0,002
Stadsbus_NOL	-208,39	39,58	-0,14	-5,27	0,000							Tram_NOL	304,70	75,75	0,15	4,02	0,000
Parking_spaces	2,53	0,57	0,11	4,45	0,000							Sprinter_Freq	205,63	30,47	0,31	6,75	0,000
Bicycle_parking	1014,51	185,71	0,16	5,46	0,000							IC_Freq	315,35	41,94	0,29	7,52	0,000
Pot_Inw	0,89	0,08	0,36	11,65	0,000												
Pot_MBO	1,08	0,37	0,07	2,91	0,004												
Pot_HBO	0,87	0,19	0,09	4,57	0,000												
Model Summary						Model Summary						Model Summary					
N	Model Sig.	R	R Square	Adjusted R Square	Std. Error	N	Model Sig.	R	R Square	Adjusted R Square	Std. Error	N	Model Sig.	R	R Square	Adjusted R Square	Std. Error
304	0.000	0,94	0,89	0,88	776,18	117	0.000	0,92	0,84	0,83	423,06	190	0.000	0,91	0,84	0,83	1007,1

4.7 GEOWEIGHTED CALIBRATION

Next step to improve the model predictions is applying geo-weighted regression to some of the models regular regression models. Geo-weighted regression will counter the problem that the model coefficients are fixed for all geographical areas. Certain variables, for example the importance of bike parking facilities, might be more important in certain specific areas. Of course, a major precondition for a good geo-weighted model is that there actually is a geographical difference in the sensitivity for certain variables. If there is no or little geographical variation, a geo-weighted model will not perform better than a regular regression model.

GWR basically performs a minor regression analysis for every station on the basis of the nearest neighbours. Separate coefficients are then estimated for each station. These separately estimated coefficients therefore can reveal regional sensitivities to certain variables. Geo-weighted calibration of the regression models is applied to the general regression models, the basic regional and, the basic main line model.

However, since the regional and main line models already are pre-selected on the basis of their location in the network, it is expected that these models will not be as much improved as the general models. The general models are not yet categorized on the basis of their location and are expected to be improved by allowing geographical flexibility in the model.

In Table 19 an overview is found with the model fit of the regression models before and after geo-weighted calibration. It appears that only the regional and main-line models are slightly improved by geo-weighted calibration. The general model fit (basic and extensive) have both slightly decreased. The extensive regional and main-line models are not geographically calibrated since this resulted in invalid models due to the relative high number of variables and low number of cases.

Table 19: Model fit of the geo-weighted and the regular regression models

GWR Model	R-square (GWR)	R-square (normal regression)
General Basic	0,838	0,840
General Ext.	0,871	0,876
Regional Basic	0,711	0,728
Main Basic	0,812	0,798

The reason for this small difference between the normal and geo-weighted models can be the fact that there is no or little geographical variation. In Map 5, which shows the coefficients of several variables of the extensive general model, it becomes clear that there is some geographic variation in the variables.

The measure for network quality Map 5 (left) has a high coefficient in the Randstad region while it has a lower value in the North, East and South of the country. This means that the connectivity of stations is more important for attracting passengers in the Randstad region (urban area with in already a high accessibility) than it is in the rest of the country (mostly rural areas with a lower accessibility).

The coefficient for the number of inhabitants Map 5 (middle) tends to be high in the Northern and Eastern parts of the country. An explanation for the difference in the size of the coefficient between the Southern Randstad area and the rest of the country is the fact that the population variables are all estimated based on survey data that originates from this Southern Randstad area. The potential that has been calculated for each station is based on distance decay curves. Also the station choice model is based on this region. This makes it likely that part of the difference in the coefficient for this variable can be ascribed to this effect.

However, the parts with a higher coefficient are generally rural areas with only a few, relative compact larger cities. Therefore it can also be assumed that the ridership at sprinter stations is for a large part

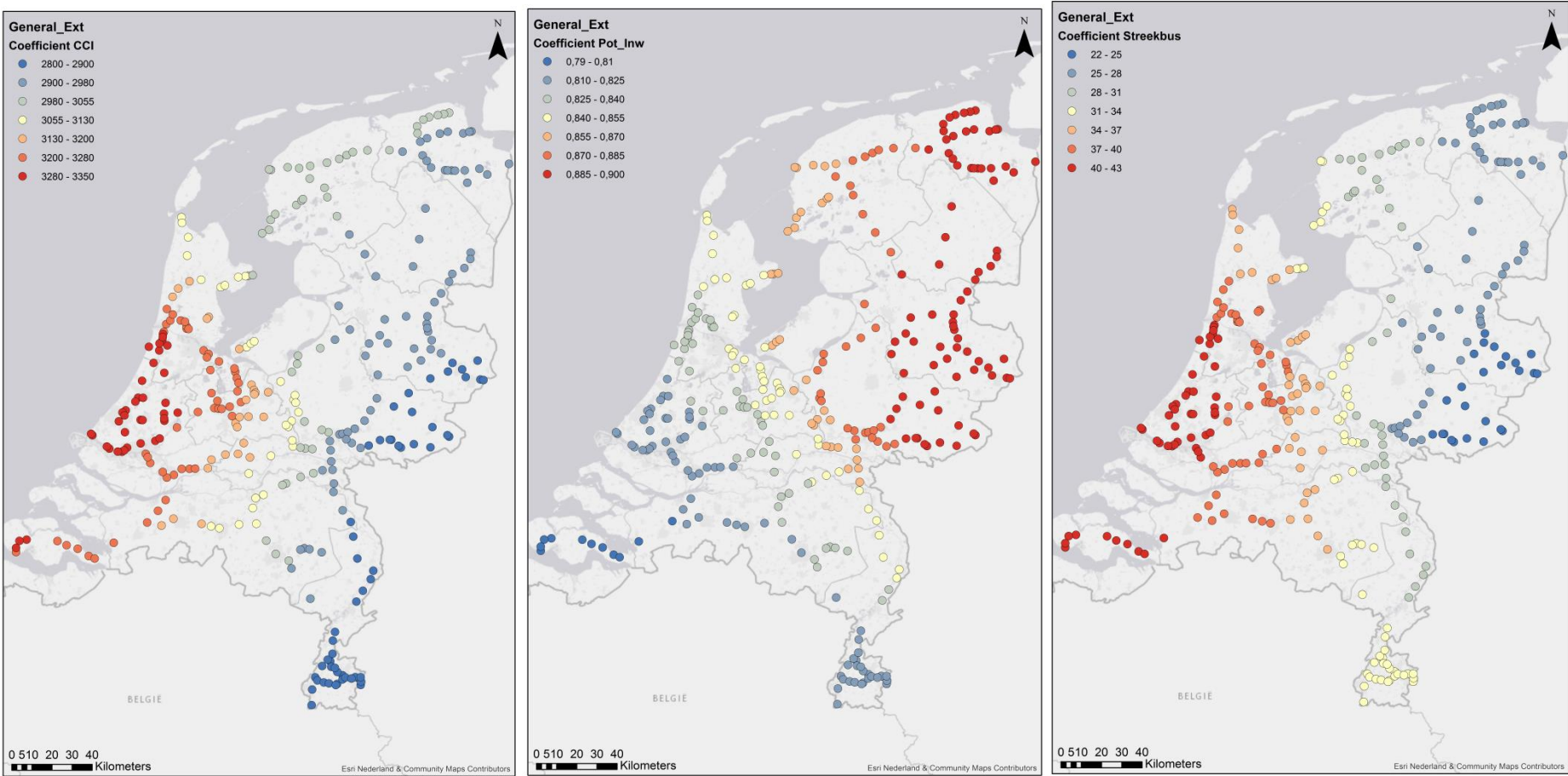
explained by the number of inhabitants without much interference from other factors. In the Randstad area and the more urban Southern areas of the country other factors such as other public transportation options, (rail) accessibility which become more explanatory for ridership as well.

An example of this effect is shown in Map 5 (right). The coefficient for the number of regional busses is high in the urban Randstad area indicating a relative larger interconnectivity between train and other modes. In the North-East this effect is less visible resulting in a lower coefficient for regional busses.

Conclusion

Geographic calibration did not improve any of the models significantly. Only a slight improvement of the basic regional and main-line models was found. The general model fit declined a bit. However, this does not necessarily means that these models should not be put into practice. All models including the regular regression models still have to be validated and have to be tested in practice. Therefore no models should be excluded in this phase.

Map 5: Coefficients for the variables CCI (left), Pot_Inw (middle) and, number of lines for regional busses (right).



4.8 MODEL VALIDATION

Regression models

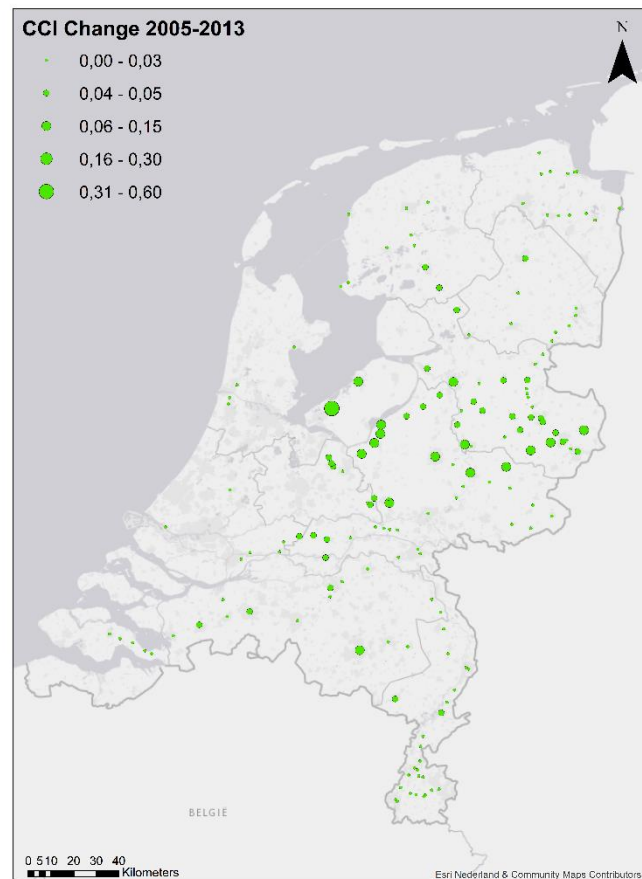
Validation of the regression models is done by applying the models on stations opened between 2005 and 2013. The year 2005 was chosen because data was still available for this year while the number of stations (cases) was large enough as well (see section 3.5 Model Validation).

As the base year is 2005, all variables included in the model will be adapted for 2005. The majority of the variables could be retrieved from the Statline data of the CBS. The CCI and SCI indicators however need to be recalculated.

The recalculation of the indicators is done by using the rail network and frequencies of 2005. This means for example that important infrastructure improvements such as the Hanzelijn are not yet available. It is therefore expected that the scores for the indicators of stations close to or along the future Hanzelijn will therefore have much lower indicator scores in 2005 compared to 2013. Also frequency changes and changes in the number of lines might be the cause for some changes in the indicator scores.

As can be seen in Map 6, the majority of the large increases in the CCI score can be found in the North-East of the country. This indicates that the opening of the Hanzelijn and the resulting redesign and frequency changes of adjacent lines have had a large impact on the overall rail accessibility.

The weighted population variables were recalculated as well with the use of population data of 2005.



Map 6: The change of the CCI indicator between 2005 and 2013 due to the opening of the Hanzelijn.

All validation stations and their corresponding model estimated are found in *Appendix 7*. In Table 20 a summary is found only containing the average model outcomes. Overall, the station estimated are close to actual demand. However, the GWR as well as the regular regression both have the tendency to overestimate demand for the more urban situated stations such as Den Haag Ypenburg, Groningen Europapark and Almere Poort.

Also for smaller stations the relative error might seem large. For the station of Gaanderen for example the estimations contain a relative error of at least 40%. The absolute however is only 146 which is less than 20% of the standard error of for example the basic regional model. This means that for larger stations the relative error is more important while for smaller stations (less than 1000 predicted) the absolute error becomes more important.

Rural stations and commuter station in smaller cities however are on average well predicted. Boven Hardinxveld, Twello, and Eygelshoven-Markt are on average correctly predicted. There is only one negative forecast for Heerlen Woonboulevard. It should be noted that this however can be interpreted as less than 100 boardings a day. Problem is that this station lacks the potential in order to overcome the constant in the regression function.

Table 20: Overview of model estimates

Name	Actual	Actual (year after)	Average Regression	Average GWR
Sliedrecht Baanhoek	553		1632	1726
Groningen Europapark	989		2191	2757
Hardinxveld Blauwe Zoom	246		493	390
Halfweg	1478	1487	2939	3197
Amsterdam Holendrecht	3176	3024	5250	5122
Gaanderen	339		533	485
Almere Poort	2256	2256	3514	4098
Apeldoorn De Maten	619		940	879
Den Haag Ypenburg	1801	908	2717	2753
Purmerend Weidevenne	1646	1644	2226	2089
Mook Molenhoek	1224		1650	1774
Boven Hardinxveld	343		461	349
Apeldoorn Osseveld	1040	640	1277	1227
Arnhem Zuid	2790		3225	3526
Helmond Brandevoort	1021	744	1169	1132
Sassenheim	3000	3000	3333	3694
Voorst-Empe	342		364	298
Amersfoort Vathorst	2559	1132	2710	2931
Twello	1554	1224	1541	1477
Eygelshoven Markt	285		272	231
Kampen Zuid	1141	1141	1055	1059
Heerlen de Kissel	371		334	299
Klarenbeek	283		251	190
Barneveld Zuid	900		774	706
Tiel Passewaaij	1269	952	1030	911
Westervoort	2250		1636	1794
Hoevelaken	1500		1036	1128
Heerlen Woonboulevard	85		-55	-71
Dronten	3142	3142	2030	2009
Maarheeze	1258	1176	669	621
Utrecht Leidsche Rijn	4700		1860	2162
Hengelo Gezondheidspark	1450		437	467

Station choice model validation

The station choice model is validated in a similar way. The choice is applied before and after the validation stations are opened. The regression model is then applied twice: One time with a potential and rail accessibility indicator as calculated before, and with a potential plus accessibility indicator as calculated after opening of the new station. Again, variables and population figures from 2005 are used as initial input.

In Table 21 the full application of the station choice model is shown. These results are demand changes as a result of the shift in station choice. Increase and decreases in the accessibility indicator are not yet included. All stations in this table are influenced by new stations opened between 2005 and 2013. This means the additional competition caused a lower demand for existing stations. Total potential has decreased for all of these existing stations.

It can be noticed that the decrease in demand as predicted by the model is accurate (with a max error of 50%) in about half of the cases. However, the other half of the stations this decrease in demand is over- or underestimated with an error larger than 50%. For a large number of stations this is because the total potential in an area is increasing, even within this one year timeframe. Demand for stations near new developments are therefore still growing while the model expected a decrease in demand.

Secondly, other factors beyond the scope of this model can be the cause for minor fluctuation in passenger demand. Demand is not the same every year. A small increase or decrease of +/- 5% is no exception, especially at the smaller stations. Attributing all change in demand to the opening of a new station is not reasonable.

The effect of opening a new station is also less clear at larger stations in Table 21 such as Hengelo, Apeldoorn and Amsterdam Bijlmer ArenA. Since the catchment area of these stations are much larger than that of the average sprinter station, there is a smaller sensitivity for changes in demand. Also

because of the large catchment area, other (spatial) developments might muddle the model results even further.

Table 21: Change of demand for stations in the vicinity of new stations opened between 2005 and 2013

Station	Station opened nearby	Potential 2013	Potential 2005	Change (abs)	Change (%)	Actual Change after year
Abcoude	<i>Muiderpoort</i>	2663	2927	-264	-9%	-29%
Heerlen	<i>Woonboulevard, de kisse</i>	16802	17626	-824	-5%	-32%
Kampen	<i>zuid</i>	3451	3992	-541	-14%	-16%
Amsterdam Bijlmer ArenA	<i>holendrecht</i>	15763	16092	-329	-2%	2%
Helmond 't Hout	<i>Brandevoort</i>	1190	1402	-213	-15%	-21%
Voorhout	<i>Sassenheim</i>	2533	2865	-333	-12%	7%
Apeldoorn	<i>de Maten, Osseveld</i>	24001	24917	-916	-4%	4%
Hoensbroek	<i>woonboulevard</i>	147	229	-82	-36%	-26%
Nijmegen Dukenburg	<i>Goffert</i>	1519	1774	-254	-14%	-14%
Purmerend	<i>Weidevenne</i>	3456	3892	-437	-11%	-6%
Tiel	<i>Passewaaij</i>	5026	5540	-514	-9%	-11%
Hengelo	<i>Gezondheidspark</i>	13662	13958	-296	-2%	-10%
Amersfoort Schothorst	<i>Vathorst</i>	5000	5332	-332	-6%	1%
Diemen	<i>science park</i>	3324	3541	-217	-6%	5%
Bunde	<i>woonboulevard</i>	668	736	-67	-9%	-7%
Nijkerk	<i>Vathorst</i>	2700	2898	-198	-7%	-4%
Nijmegen	<i>Nijmegen Goffert</i>	33762	34350	-588	-2%	0%
Helmond	<i>Brandevoort</i>	6457	6818	-361	-5%	19%
Purmerend Overwhere	<i>Weidevenne</i>	3969	4240	-271	-6%	-16%
Elst	<i>Arnhem Zuid</i>	3913	4038	-125	-3%	-5%
Duivendrecht	<i>Holendrecht</i>	20762	20815	-54	0%	-2%
Almere Muziekwijk	<i>Poort</i>	4867	5055	-188	-4%	-6%

Stations where the choice model is relative successful are sprinter stations such as Nijmegen Dukenburg, Almere Muziekwijk and Bunde which are locally oriented sprinter stations near existing residential developments. Also these stations are less dependent on passengers arriving by public transport or car. Sassenheim for example is a station which is more transit oriented with extensive park & ride facilities. This makes it more difficult to estimate the impact of a station on existing stations.

Effect of rail accessibility indicator

When the effect of the rail accessibility indicator is taken into account as well, the results will slightly change. Since the connectivity of a station determines part of the ridership level as well, the effect of a changing CCI indicator after opening a new station is measured as well.

For the same set of stations in the situation before and after the new station was opened, the CCI values are calculated. Again the regression functions are used to make a new estimation. In order to isolate the effect of the change in the indicator the potential was kept the same with base year 2005.

Table 22: The demand change at existing stations near new stations as a result of changes in the CCI indicator values.

Station	Actual (2013)	Average model estimation (with data 2005)	Demand Change (abs)	Demand Change (%)
Abcoude	1625	2909	-92	-3%
Heerlen	12374	18635	-41	0%
Kampen	4256	4025	-5	0%
Amsterdam Bijlmer ArenA	18961	16383	-140	-1%
Helmond 't Hout	1247	1413	-25	-2%
Voorhout	3452	2888	-8	0%
Apeldoorn	14015	26381	14	0%
Hoensbroek	196	221	-7	0%
Nijmegen Dukenburg	2151	1787	-29	-2%
Purmerend	2992	3935	-44	-1%
Tiel	4128	5581	-32	-1%
Hengelo	14008	14476	-16	0%
Amersfoort Schothorst	5642	5354	-11	0%
Diemen	3423	3553	-107	-3%
Bunde	954	728	-6	-1%
Nijkerk	3650	2909	-14	0%
Nijmegen	44051	35171	16	0%
Helmond	6847	6851	-6	0%
Purmerend Overwhere	2312	4269	-27	-1%
Elst	3863	4041	-28	-1%
Duivendrecht	13068	20876	-98	0%
Almere Muziekwijk	7030	5070	-96	-2%

Changes as a result of the decrease in the CCI value are in general small (Table 22). The maximum estimated demand change is a decrease of 3% for the stations of Diemen and Abcoude. These two stations are influenced by the fact that these stations have high weighting travel relations with Amsterdam central station and Utrecht central station. Along the route to both of these stations, new sprinter stations have opened. This directly increases travel time by at least 4 minutes.

For most other station these effects are smaller since they have a more balanced set of travel relations or because they are less affected by the travel time loss since the station in question also offers intercity connections which have not been affected.

In some occasions a slight increase in the CCI value is measured. This might indicate that a new station actually has an added value for the existing station since it now offers a valuable new travel direction. This can be observed at the station of Helmond for example where the existence of intercity links limits the negative travel time results of the new sprinter station of Helmond 't Hout, while at the same time this new station offers a new connection.

Conclusion

The Regression models are in general able to make accurate predictions for the validation stations subset based on 2005 data. In total 10 different models have been applied and tested including 6 regular regression models and 4 geo-weighted regression models.

Also the station choice model was tested. It appeared that the effects of a new station are best estimated for locally oriented sprinter stations in existing residential areas. When stations have larger catchment areas or are located near new (residential) developments, this effect is deluded by other effects such as the uncertain increase of demand by the increase of the population.

When The CCI indicated changes are taken into account it appears that this effect is only small with a maximum decrease of 3% measured in the validation dataset. If such a relative large decrease is measured it means that the travel time on a critical link is significantly higher because of the dwell time at the new station.

Table 23 shows the final demand change as measured by the station choice model and accessibility indicator. The final estimated demand change is, when compared with actual figures, not accurate since over- or underestimations of more than 50% are common.

Table 23: Total demand change modelled (potential and CCI combined)

Station	Total demand change modelled	Actual demand change year after opening new station
Abcoude	-12%	-29%
Heerlen	-5%	-32%
Kampen	-13%	-16%
Amsterdam Bijlmer ArenA	-3%	2%
Helmond 't Hout	-17%	-21%
Voorhout	-12%	7%
Apeldoorn	-3%	4%
Hoensbroek	-37%	-26%
Nijmegen Dukenburg	-16%	-14%
Purmerend	-12%	-6%
Tiel	-10%	-11%
Hengelo	-2%	-10%
Amersfoort Schothorst	-6%	1%
Diemen	-9%	5%
Bunde	-10%	-7%
Nijkerk	-7%	-4%
Nijmegen	-2%	0%
Helmond	-5%	19%
Purmerend Overwhere	-7%	-16%
Elst	-4%	-5%
Duivendrecht	-1%	-2%
Almere Muziekwijk	-6%	-6%

What is missing at this point is a better insight on how well each individual regression model is performing and how this whole methodology can be put into practice. In the next section the actual accuracy of the models will be evaluated and the model will be put into practice for current station proposals.

4.9 RELIABILITY OF RESULTS

It is clear that the when the regression models can produce accurate results which are close to actual ridership. However, only one value was given as a forecast for each station. This single value does not give any information on how accurate this forecast is, and with what kind of margins this forecast should be taken. This section will therefore further asses the accuracy of all regression models.

Absolute and relative error

Main point in how the accuracy of a model will be assed depends on whether the absolute or the relative error is taken. When dealing with small ridership numbers, the model error in terms of relative error (error margin in percentage) can be very large. In absolute numbers this error is fairly small. It is the other way around for stations with a large ridership where the relative error is relative low and the absolute error is high.

In Figure 17 and 18, these absolute and relative errors are plotted against the percentage of cases that falls within a certain error margin. It holds for roughly 80% of all cases that the ridership can be estimated within 1000 daily passengers' error margin. In relative terms 80% of all cases are predicted within an error margin of 50% of the actual number of passengers.

Problem is that in both figures the size of the error is exaggerated. In figure 17 the 20% of the stations with an error larger than 1000 are stations with a total ridership estimation which is a multitude of that.

Relative error for these station might therefore be minimal. In figure 18 the lowest scoring 20% are mainly small stations with only a few hundred daily estimated passengers. A relative error over 50% is not uncommon for these stations.

Comparison of models

Besides the size of the error, figures 17 and 18 are also giving information of the quality of the different models and whether these models are generally over- or underestimating demand. In total 7 regression models (4 regular and 3 geo-weighted) are included in the figures. On top of that also the input variable “total potential” is included.

The specific model type is a combination of the regional and main-line regression models. Since all stations are either fitted for the regional or the main-line model, these models will be regarded as one model containing two specific sub-models.

This total potential is the worst scoring model in absolute and relative terms. In 60% of the cases demand is underestimated with a large error margin. This means that estimating the regression functions using total potential as an input variable was of added value since all regression models are performing better.

Secondly it is noticeable that the geo-weighted models are significantly better performing than the regular regression models. In previous sections it is already discussed that the increase in the model fit is only small after geo-weighting the models. Figures 15 and 16 show that also the overall accuracy of these models is not improved nor worsened

Figure 17: Absolute error plotted against the percentage of cases that falls within the error margin

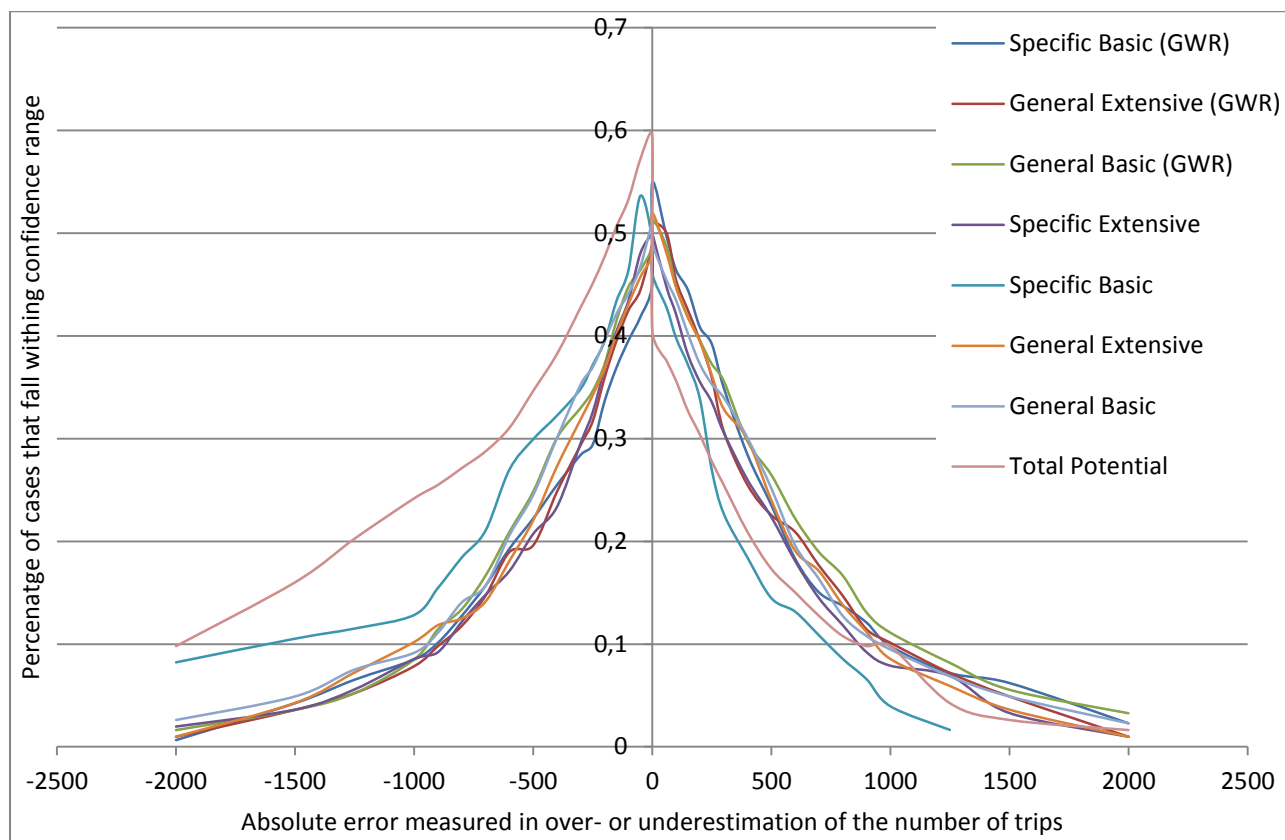
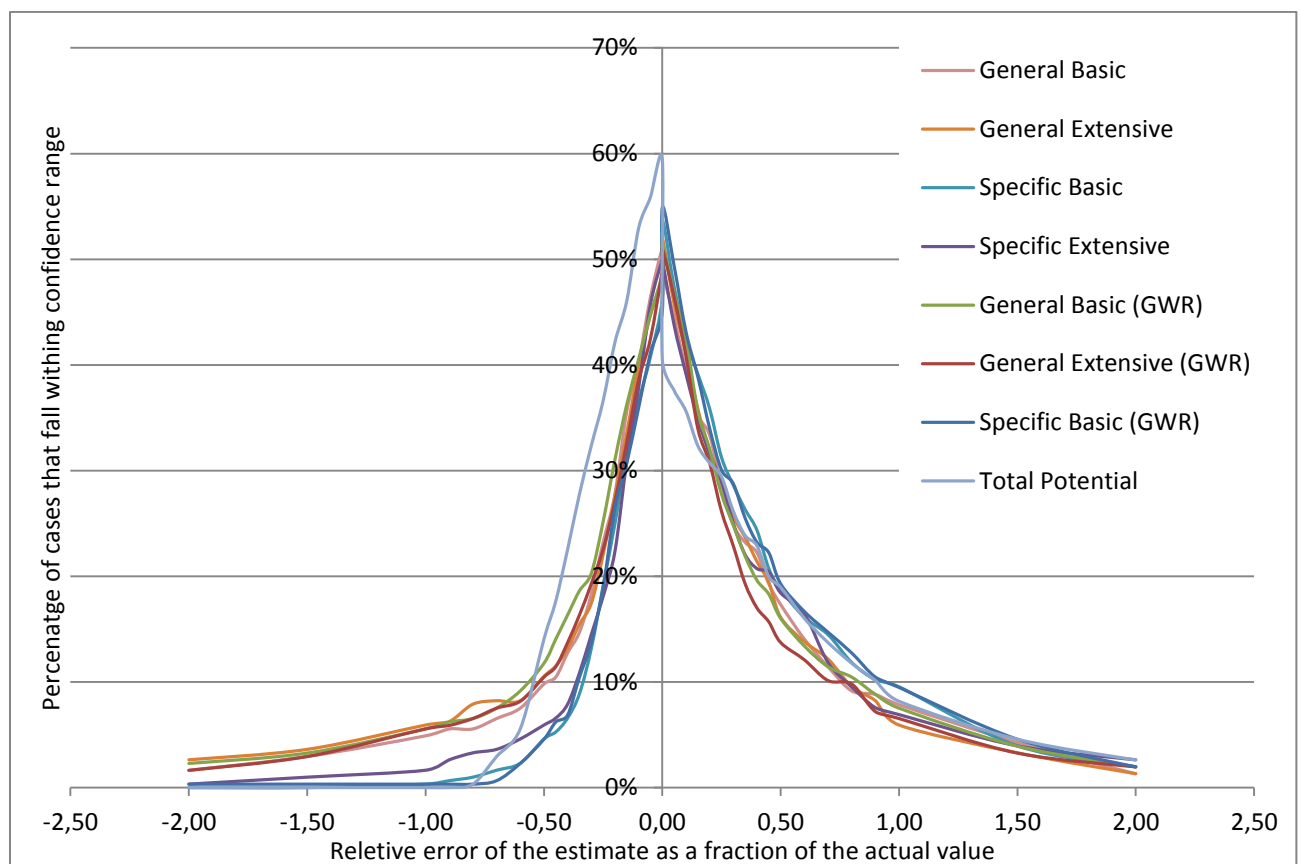


Figure 18: Relative estimation error plotted against the percentage of cases that falls within the error margin



As for the other regression models it is harder to make a clear distinction between the quality of the individual models. It depends on the type of station to determine the best model. In case it is a station with on average more than a 1000 estimated passengers a day, more weight should be given to a model that is performing well in relative terms in the top 50% of the cases. Models such as the specific extensive and the general extensive models are in that case the better choice. When the station is expected to be receiving less than 1000 daily passengers. The better choice is a model that performs well in absolute terms. The general basic model would perform better in that case.

However, the differences between the models are too small to select only one model that should be used. Based on over- and underestimation margins and error margins a method should be derived to give a final, most likely number of passengers, together with an error margin based on the outcomes of the various models.

Ideally, the outcome of this model would be compared with the same results from other demand estimation models such as the PINO model from Dutch railways. This is not possible due to the fact that data on all stations is unavailable.

Aggregating model estimations

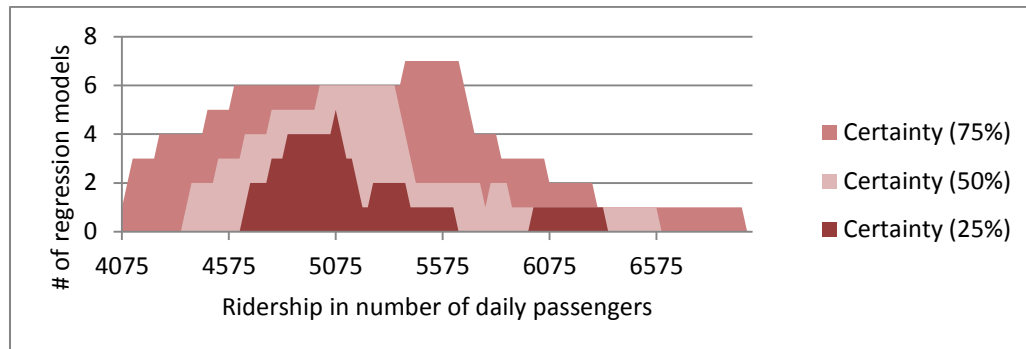
When all the regression models are applied to a station, based on figures 17 and 18, the maximum and minimum error within a certain margin is known. For example, the general basic model has a 50% chance the estimated value for a station will have an error between 0.75 and 1.30 times the estimated value. This figure is based on the estimations done for all stations during the model calibration.

When for example three margin errors (25, 50 and 75% certainty) are selected it is possible to indicate the margins within an estimation is valid. This is graphically depicted in figure 19 for the station of

Meppel. Vertically depicted are the number of models of which the corresponding ridership estimation on the horizontal axis is within their error margin.

Cumulatively, also the percentage of certainty is depicted. The highest chance is that the actual ridership will be located close to areas in the graph within the 25% score area. When moving to 50% or 75% areas, the chance that the right ridership figure is within this margin becomes larger, but the error margin becomes larger as well.

Figure 19: Relative Error margins for Station of Meppel with a 25, 50 and 75 percent certainty.

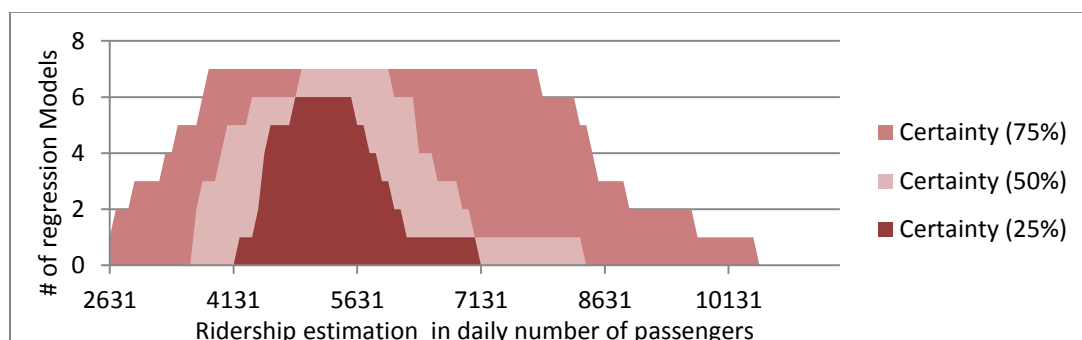


Whereas individual models predicted ridership figures for Meppel ranged from 4874 to 6176 this method has reduced this uncertainty to 5075 as the most likely number of passengers within a certain error margin. Actual ridership for this station is 5346, which is within the 25% certainty margin.

However, within this method it is also important to pay attention on the error type used: relative or absolute. Because in the previous example a relative large sprinter station was chosen with a couple of thousands of daily boardings, the relative error margin works well. When the margin errors are translated in terms of absolute error margins it works out differently (Figure 20).

Since larger stations generally have a larger absolute error, the overall margin error becomes much larger. The final number of passengers predicted according to this method with the use of absolute margin errors is 5219. Although this is in this case closer to the actual ridership figure of 5346, the accompanying error margin is much larger. In case this was a proposed station of which the actual ridership is not (yet) known use of the relative error margins is therefore advised. For smaller stations with less than 1000 passengers a day the use of absolute error is more suitable.

Figure 20: Absolute Error margins for Station of Meppel with a 25, 50 and 75 percent certainty.



Conclusion

This section has presented the quality and accuracy of all regression models. It is demonstrated that the use of a regression model has an added value as estimation results are better with regression models than when only the total potential is used. However, difference between the regression models is small making it not possible to select a best model.

A solution was found in combining the models and use the error margins of individual models to come up with a final aggregated station score. This give a most likely number of passengers together with an error margin.

4.10 MODEL APPLICATION

Ridership

The models are applied to some of the proposed stations as described in Appendix 1: proposed stations in the Netherlands. Table 24 gives an overview of all of these stations together with an aggregated ridership estimation based on relative and absolute error margins as discussed in the previous section.

Table 24: List of proposed stations with an estimation based on absolute and relative error margins

In general these figures are all positive. The only exceptions are the proposed stations in the Eemshaven, Wildervank, Leeuwarden-Werpsterhoek and Sneek-Harinxmaland. The station in the Eemshaven can be considered an exception in many ways since the very low proposed frequency (less than once every two hours) and its dependency on the ferry to the island of Borkum makes this a station of which the ridership that cannot be estimated with the use of this model.

As for the station of Wildervank the negative forecast is simply the result of a too low demand. In an actual situation ridership cannot be negative and some ridership can be expected. However, according to this model this ridership is not enough (less than 100 passengers a day) for a reasonable forecast.

The latter two stations have negative forecasts most likely because these are stations in new greenfield developments. For these stations a rough estimation for the number of inhabitants was made based on the number of dwellings that is to be built. However, often infrastructure is not yet in place and the data on the number of new residents in not correct or incomplete. This

results in forecasts that are not entirely reliable. However, more reliable forecasts would be possible for these stations if the input data (network dataset, population) is adjusted with the new (proposed)

Station	Relative error margin	Absolute error margin
's-Hertogenbosch Maaspoort	1483	1471
's-Hertogenbosch Avenue	1468	1455
Apeldoorn West	2118	2090
Arnhems Buiten	919	884
Baexem	825	776
Belfeld	454	411
Berkel Enschoot	1807	1844
Breda Oost	2383	2345
Deventer Platvoet	1239	1228
Deventer Zuid	1458	1446
Duurkenakker	155	61
Eemshaven	n.a.	-948
Eindhoven Airport	1996	1963
Geldermalsen Zuid	1003	1045
Gorinchem Noord	287	350
Haelen	766	723
Hazerswoude Koudekerk	1762	1740
Hoogkerk	494	397
Leerdam Broekgraaf	271	133
Leeuwarden-Werpsterhoek	116	-41
Lelystad-Zuid	770	663
Maartensdijk	2594	2635
Nijkerk Corlaer	1287	1314
Oss Oost	903	869
Ressen	945	849
Schiedam Kethel	3561	3739
Sneek-Harinxmaland	29	-88
Stadskanaal (centrum)	1574	1613
Staphorst	827	835
Stroe	850	829
Utrecht Lage Weide	3678	3900
Utrecht Majella	2223	2408
Utrecht Vaartsche Rijn	2768	3003
Venlo Grubbenvorst	404	359
Wijchen Oost	597	697
Wijchen West	845	798
Wildervank	33	-16
Zevenaar Oost	305	182
Zoeterwoude Meerburg	1372	498
Zwolle Stadshagen	366	233
Zwolle Zuid	942	901

developments. This data however, is not always available. Finally a selection of proposed stations can be found in Appendix 9 with error margin graphs as well.

Abstraction and rail accessibility effects

Besides the total ridership of a new station, it was identified in the theoretical framework section 2.2, that a new station also abstracts demand from existing stations and will decrease the overall rail accessibility of existing stations. In Table 25 an overview of all stations is found that are directly affected by the opening of one of the new stations in table 20.

Table 25: Overview of stations affected by the opening of a new station. This includes the decrease in demand by abstraction and as an effect of a reduced rail accessibility.

Station	Demand change due to CCI (%)	Demand Change due to abstraction (%)	Demand Change total	Station	Demand change due to CCI (%)	Demand Change due to abstraction (%)	Demand Change total
Sneek	0%	-6%	-6%	Alphen aan den Rijn	0%	0%	0%
Sneek Noord	0%	1%	1%	Schiedam Nieuwland	-1%	0%	0%
Leeuwarden	0%	-3%	-3%	Schiedam Centrum	0%	-2%	-2%
Deinum	-1%	-3%	-4%	Arkel	-2%	-23%	-25%
Mantgum	-1%	-1%	-2%	Gorinchem	-1%	-11%	-12%
Zuidhorn	0%	0%	0%	Leerdam	-3%	-63%	-66%
Groningen	0%	-4%	-4%	Geldermalsen	-1%	-9%	-9%
Roodeschool	12%	-1%	11%	Breda	0%	-23%	-23%
Veendam	1%	-12%	-11%	Oisterwijk	0%	-4%	-4%
Zuidbroek	0%	-9%	-9%	Tilburg	0%	-4%	-4%
Scheemda	0%	-1%	-1%	Eindhoven Beukenlaan	-1%	-6%	-7%
Zuidbroek	0%	-9%	-9%	Eindhoven	0%	-5%	-5%
Scheemda	0%	-1%	-1%	Best	0%	0%	-1%
Meppel	-1%	0%	-1%	Hertogenbosch 's Oost	-2%	-18%	-20%
Zwolle	0%	-5%	-5%	Hertogenbosch 's	0%	-8%	-8%
Kampen	0%	0%	0%	Rosmalen	-1%	-26%	-28%
Kampen Zuid	-1%	0%	-1%	Oss West	-5%	-2%	-7%
Zwolle	0%	-5%	-5%	Oss	0%	-3%	-3%
Wezep	-1%	0%	-1%	Wijchen	-1%	-27%	-28%
Dalfsen	0%	0%	0%	Ravenstein	-6%	-4%	-10%
Lelystad Centrum	0%	-10%	-11%	Nijmegen Dukenburg	-2%	-8%	-10%
Oldenzaal	0%	-4%	-4%	Elst	0%	-2%	-3%
Apeldoorn	0%	-38%	-38%	Nijmegen Lent	-5%	-7%	-12%
Apeldoorn Osseveld	-1%	-6%	-7%	Oosterbeek	-1%	-24%	-26%
Apeldoorn De Maten	0%	-9%	-10%	Arnhem Zuid	-1%	-1%	-2%
Twello	-1%	-3%	-4%	Arnhem	0%	-5%	-5%
Deventer	0%	-33%	-33%	Arnhem Velperpoort	0%	-1%	-1%
Deventer Colmschate	0%	-36%	-36%	Zevenaar	0%	-5%	-5%
Barneveld Noord	0%	-1%	-2%	Didam	-1%	-8%	-9%
Nijkerk	-1%	-12%	-13%	Swalmen	-1%	-1%	-2%
Amersfoort Vathorst	0%	-6%	-7%	Roermond	0%	-3%	-3%
Hollandsche Rading	-16%	-31%	-47%	Tegelen	-1%	-30%	-30%
Maarssen	0%	-1%	-2%	Reuver	-1%	-5%	-6%
Utrecht Leidsche Rijn	0%	6%	6%	Blerick	-1%	-6%	-7%
Utrecht Zuilen	0%	2%	2%	Venlo	0%	-8%	-8%
Utrecht Centraal	0%	-3%	-3%	Horst-Sevenum	0%	-1%	-1%
Utrecht Lunetten	3%	0%	3%	Leiden Lammenschans	0%	-3%	-3%
Utrecht Overvecht	0%	0%	-1%				

In general the effects of a reduced rail accessibility are not large. In most cases this will result in a demand reduction of only 1 or 2%. There are however exceptions. In the case of the station Hollandsche Rading a large decrease in accessibility is also estimated to have a large effect on the total ridership of this station. In the situation before the new station Maartensdijk was opened, this sprinter station was ideally located between Hilversum and Utrecht. It had a very high accessibility score.

As for the stations along the Merwedelingelijn (Leerdam, Arkel) and the IJsellijn (Nijmegen-Lent, Ravenstein, Nijmegen-Dukenburg) it is a combination of multiple new stations that reduces the accessibility above average.

The accessibility of Roodeschool and Veendam increases after opening of consecutively station Eemshaven and Stadskanaal. These stations do not suffer any negative effects since they are currently end of the line. The new stations are extensions of these lines whereas Veendam and Roodeschool only have on additional connection without any travel time loss.

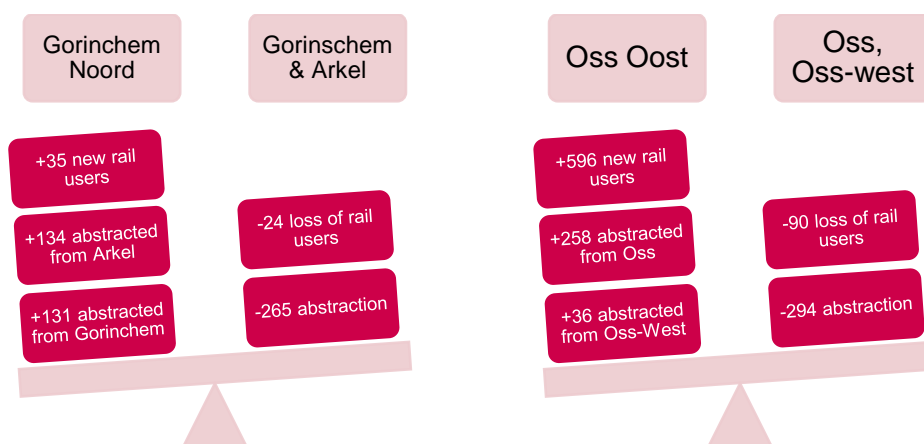
Final station balance

The final question is now when it is feasible to build a new station (keeping infrastructural limitations are kept out of the equation). At this point a ridership estimation is made for many potential stations in the Netherlands. At the same the effects of these new stations have been estimated as well. Bringing these two factors together can give more information on how a new station performs in for example attracting new passengers.

Gorinchem for example currently receives 1190 passengers a day, Arkel 581. The new proposed station of Gorinchem-Noord is estimated to receive around 300 passengers a day. As an effect of Gorinchem-Noord it is expected that because of the reduced rail actability Arkel loses 2% of total ridership, Gorinchem 1%. Demand abstraction at Arkel is estimated to be 23% at Gorinchem 11%

Too summarize all effects with the re-use of figure 1 from section 2.2 see the example stations of Gorinchem-Noord and Oss-Oost in Figure 21:

Figure 21: Total balance for the stations of Gorinchem Noord (left) and Oss Oost (right)



In the case of Gorinchem-Noord, the balance is slightly tilting to the left in favour of the new station since the number of rail users in this case will slightly increase. However, when other factors are taken into account (investment costs, operation costs) this balance will almost certain be tilting to the right. Also from the viewpoint of the national government a station needs to receive at least 1000 passengers a day before it can be feasible.

In the case of Oss-Oost the feasibility is higher. In total 596 new users are attracted against a loss of 90 passengers due to the reduced accessibility at mainly the sprinter station of Oss-west. The intercity status of the station of Oss is also the reason why demand abstraction is less of a problem in this case. Oss is also hardly affected by longer travel times because of this intercity status.

5. DISCUSSION

At this point a ridership estimation can be made with the use of a station choice model, rail accessibility indicator and several regression models. Finally also the accuracy of these models is assessed and based on that a method was derived to aggregate model results into a final ridership estimation.

This section will discuss the separate steps until the final ridership estimations by looking back to literature as discussed in the literature review. How do these ridership estimation models hold in contrast to other models which have been developed?

5.1 THE USE OF THE RAIL ACCESSABILITY INDICATOR

The use of a rail accessibility indicator or “rail service quality indicator” (RSQI) as described by (Debrezion, et al., 2009) proved to be a valuable addition to station choice models in previous researches. Also in this model such an indicator was useful.

However, unlike in the research of Debrezion et al. (2009), the closeness centrality index (CCI) is not only used in the station choice model but is also used in the regression analysis. Therefore this variable is in this research used as a measure to distribute passengers over stations with the choice model on a local level, but also as a measure for overall rail attractiveness on a national scale.

Since similar variables has not been used in regression models estimating rail ridership, a comparison of the use of this exact variable with literature is not possible. However, many similar variables have been used, often with similar success. The main difference is that in most cases accessibility is something that was viewed from one central location. In other words, the variable was defined as the accessibility to for example the city centre of a large city.

Blainey et al. (2010) used the travel time to the centre of Cardiff, Wales as a measure for accessibility. In this thesis the variable CCI was able to be defined not as a measure of accessibility to one central point, but to the rest of the network as well by weighting the various links based on the gravity model.

However, this indicator could not replace all other rail network related variables. Especially the frequency, which is an important factor in rail accessibility, could not be entirely replaced by the CCI. Though some correlation exists between the variables it was not high enough to exclude one of the two variables. In a way they seem to supplement each other indicating that frequency is not only a measure of accessibility, but also for example for comfort or ease of travelling.

Finally, it should be noted that initially there were two accessibility variables. Besides the CCI there was also the Straightness centrality index (SCI). This variable was meant to be an indicator for attractiveness of train travel opposed to travel by car. In later stages in this research, when estimating the choice model and regression models this variable (which was correlating with the CCI) was always of lesser importance. It is therefore not used in any model.

5.2 STATION POTENTIAL & STATION CHOICE MODEL

The use of distance decay curves and the station choice model did provide for good population variables that are used as input for the regression. In literature (Gutiérrez, et al., 2011) however, this method did not yet include an advanced method to deal with competition between stations. Therefore in this thesis the station choice model was used to add a third dimension in dealing with catchment areas. By also including the number of jobs and students in higher education several ‘potential’ variables are estimated.

In regression models this total potential proved to have a better explanatory value compared to regular population variables. A potential variable is used in all regression models making it together with the CCI a base variable, essential in explaining ridership.

The station model (model 3) that is chosen to be used was selected because the model is intuitively satisfying since demand at stations can only decrease after a new station is included. This was achieved by only estimating two alternative specific constants to remove any implied perceptual rank-based difference as was done by Blainey and Evans (2011) as well. The problem with model 1 & 2 is that these models still do contain an ASC for every choice option making the models conceptually unsatisfying regardless of the better model fit.

The coefficients estimated in the station choice model 3 for the distance to intercity train stations are very similar to the coefficients for cycling (-0.0008), the distance to sprinter train stations resembles the coefficient for walking (-0.0012) as based on literature (Givoni & Rietveld, 2014).

5.3 REGRESSION MODELS

Variables used

In total six regression models have been estimated. By estimating different regression models applicable in different situations the aim was to increase the accuracy of the estimations. Whereas it was also expected that different variables were explanatory for the different models this was however not the case.

All models are populated with (sub-) variables of the potential, network quality and transferability to other modes. Variables related to socio-economic circumstances (income, employment, and car-ownership) or other spatial features (land-use, land-use mix, design) all returned no further significance in the regression models. In literature variables such as income (Blainey & Mulley, 2013), station design/architecture (Cascetta & Cartení, 2014), Car ownership (Wardman, et al., 2007) did have an influence on ridership as well, but in this research this could not be confirmed.

There are two exceptions however. First of all the variable “proximity”, which measures the average distance to several urban services, did have explanatory value in the regional models whereas the CCI indicator was of lesser importance. Likely reason for this is the more singular focus point that exists on these regional lines. Instead of multiple possible destination that are important, regional lines are usually more focussed on only one or two main attracting destinations. The variable “proximity” fulfilled the same role as for example the variable “distance to CBD” (Liu, et al., 2013) or “distance to Cardiff city centre” (Blainey & Preston, 2010) which works well in regions with only one main attracting destination.

Secondly the proximity variable might catch the effect of a general higher public transport usage when the distance becomes larger. Especially when the distance becomes too large for cycling, the train might be the only option to travel by public transport to the nearest urban centre with services such as education. Bus lines on the same route are often not available and thus competition from other public transportation modes is non-existent.

The other variables that are breaking the habit of reusing the same variables are the number of parking spaces and the availability of guarded bicycle parking. These two variables do have an explanatory value in the extensive basic model as also was demonstrated in literature (Blainey & Mulley, 2013). In the main-line and regional model version they are no longer significant. This mainly has to do by the fact that there are only a few stations significantly dependent on park & ride facilities. Guarded bicycle facilities are only available at the larger stations. Using these variables in a relative small sample group, these variables have little explanatory value.

However, it does not mean that variables that did not have an explanatory value in this research do not have any influence on rail ridership. But in this research rail ridership was estimated on a national

level where variations in these variables tend to be more subtle. Other demand estimation researches were focussing on metro or light rail systems in which these subtle variation in certain variables might be more important.

Regression coefficients

Although the same (sub-) variables were used in multiple models, the coefficients of these variable did differ. The coefficient of potential variables did score higher in the regional models which consist of stations in more rural areas. At the same time the coefficients for number of other public transportation lines and for the CCI indicator are smaller. This indicates that rail ridership in rural areas is mainly determined by its potential and less dependent on other factors such as accessibility and transferability.

At the same time is ridership in urban areas more dependent on other factors. Here for example, also the type of transferable public transport is of importance. City busses often act as competing modes while light rail/tram acts as a feeder. Also the type of service depends on ridership. A higher frequency of intercity trains leads to a higher ridership than the same frequency of sprinter trains.

Use of geo-weighted regression

The use of geo-weighted calibration after the regression models are estimated did not meet expectations since it did not improve the model fit significantly of any of the regression models. Secondly because also the GWR models did not have an improved accuracy compared to the existing regression models.

This outcome is contradicting with other researches (Blainey & Mulley, 2013), (Blainey & Preston, 2010) among others who did found an increased model fit after application of geo-weighted regression. However, an important condition for a successful GWR is that there actually is geographic variation in the variables researched. The variables researched in this thesis might not have been prone to geographic variation or maybe they were already geographically adjusted (CCI indicator). However, certain variables such as the number of bus lines, and the CCI indicator did show some geographic variation after application of GWR which suggest that some geographic variation is present.

This however, does not mean the application of geo-weighted regression was to no avail. It also has shown that a variable such as the total potential of a station does not differ much around the country. Although the coefficient of the total potential is higher in rural parts, there is no indication that the station choice model and distance decay curves are prone to error because it was calibrated using data from the South Holland region only.

6. CONCLUSIONS

6.1 RESEARCH QUESTIONS

In this research the following aspects have been successfully adopted in order to estimate a rail ridership demand estimation model:

1. A rail accessibility index (CCI) that estimates an index for any station in the Netherlands on the basis of the accessibility by train in terms of in vehicle travel time and number of transfers.
2. Distance decay curves based on station type. On the access side, eight curves have been estimated: one for each station type plus one for intercity and for sprinter stations. On the egress side three curves have been estimated: for sprinter, intercity and combined.
3. A multinomial station choice model. Based on a choice set of two intercity and two sprinter stations the probabilities of each station can be calculated per six digit postcode area. Based on this model the potential of train users from the number of jobs, population and student places can be derived. The effects of a new station can be measured as well.
4. A regression analysis with 6 regular regression models and 4 geo-weighted regression models.

Station specific variables in relation to the catchment area

To answer sub-research question five *"How do station specific variables (such as station type, - quality, and – facilities) impact the station catchment area?"*:

Station specific variables can have an influence on the catchment area. The distance decay function that were calibrated on the basis of station type show a large difference between the size and trip generation of the catchment areas of intercity stations and sprinter stations. An Intercity station catchment area can be 15 kilometres wide. Sprinter stations however will only have an effect till 5 kilometres.

Also within the group of intercity and sprinter stations differences are observed. Type 5 (suburban stations) tend to have the smallest catchment area while type 4 sprinter stations have the largest catchment areas amongst the sprinter stations. Type 1 intercity stations have the largest catchment areas of all stations; type 2 intercity stations have smaller catchment areas.

Addition to the size of the catchment area is the trip generation which is also different for the various station types. Especially on a short distance from the station, type 1 and 2 stations attract much more rail passengers than a type 4, 5 at the same distance. The results of the type 6 distance decay curves are however less useful. Because of the low number of observations it was not possible to estimate a reliable distance decay curve.

The catchment area can also be partly determined by the station specific variables used in the station choice model. This includes the availability of guarded bicycle parking and the number of bus, tram or metro lines passing through the station.

The effect of network specific variables

Several station specific variables were included in the models. Number of bus lines passing through the station, frequency of trains and, most important, the accessibility indicator (CCI). To answer sub question 6: *"How will network specific variables (such as reliability, accessibility and service level) influencing passenger demand at train stations?"*:

The rail accessibility index proved to be extremely useful in all further aspects in this study. Especially the closeness centrality indicator (CCI) was able to explain a large portion of the rail ridership and station choice. Also for new stations this indicator is a good explanatory variable as it is possible to calculate this index with the use of Omnitrans.

Secondly, for almost all station types it was possible to estimate logical distance decay curves with use of the Stedenbaan data. This was done for as well as the access as the egress side of the trip. Only exception were the type 6 stations because of insufficient cases the resulting distance decay curve is not reliable.

Intercity stations (types 1 & 2) have the largest catchment areas based on these distance decay functions and the highest trip generation rate. However, the type 1 stations outperform the type 2 stations on both aspects. As for the sprinter stations, type 4 stations have the largest catchment areas. However, this catchment area is half of that of an intercity station in terms of trip generation and distance. Type 5 stations have the smallest catchment area, most likely because the local character of these stations and the lack of access mode accessibility.

Competition between stations

Competition between stations, is an important factor that can determine a large part of rail ridership. The main question for this aspect is:

"How is competition between stations included and how is this influencing the total ridership demand"

Competition is included by distributing the derived demand for rail transport over all stations on the basis of distance decay curves with the use of a multinomial station choice model. This choice model differentiates between the two main competing station types: Intercity stations and sprinter station.

Variables such as the availability of guarded bicycle parking, number of bus, tram or metro lines, CCI index, and distance to the station will all influence the utility of a station. In the practical application of this model it means that in a regular situation the closest station is chosen. However, when another, more distant located station is available with much better score on one of these variables, this more distant station can be chosen as well.

Explanatory power and model accuracy

After several models were estimated the final question to be answered is:

"What is the explanatory power of the model in predicting future travel demand?"

Depending on the (assumed) size of the stations the explanatory power of the models vary. In case a small station (less than 1000 passengers a day) is being researched, more emphasis should be put into the absolute error a regression model can give. For larger stations (>1000 passengers a day) the relative error is more important.

Based on this assumption in theory some regression models will perform better for small stations than for large stations and vice versa. However, in this research the difference between the different regression models was small in absolute as well in relative terms.

A solution was found in assessing the minimum and maximum value in three (25, 50 and 75%) confidence margins. Based on this method a ridership estimation can be given with not only a single figure but also with a margin.

Final ridership estimation model

Main question for this research was:

How can the daily number of passengers of a new train station be forecasted based on station choice and network accessibility?

Ridership in this model is estimated in three steps:

First a potential ridership is estimated, secondly this potential is distributed over all stations with the station choice model, and finally a regression model will give a final estimation. Following this process, this model takes into account the distance decay effect, the preference for a station based on station specific variables, and factors that influence rail demand on a national scale using the regression model.

Network accessibility is an important factor in the second and final step of this process. In the second step it can determine the utility of a station, in the third step it is a measure for overall rail accessibility on a national level.

Limitations

A large limitation of this study is the fact that it only takes into account the mode of rail transport. Competition/feeder effects are only included in the regression as separate variables but not in the station choice model. However, the competition between bus services along the same route as a train service can lead to a significant decrease in ridership.

Secondly, for the type specific regression models two reference classes have been made (regional and main-line). However, there is no classification possible in which a clear distinction between the stations can be made. At the same time there are many ways to classify the stations. Therefore there is always room for error by making a wrong classification. This could lead to an error in the final model as well.

Furthermore, this model was calibrated on data of 2013. However, the number of rail passengers has been steadily growing in the last couple of years. This results in changing numbers of passengers near not only new stations but at existing stations well. This growth in passengers will cause an increase in the overall rail trip generation. Therefore the distance decay curves; choice model and regression models will lose explanatory power when the model is applied in future scenarios.

Suggestions for further research

First of all, in this research only distance decay functions on the basis of distance were made. Since the number of observations with a known access and egress mode was limited, estimating a distance decay curve on the basis of access modes was not possible. However from the observations that were available there was a strong link between distance and mode. Related to this, a (nested) station choice model could be improved with the addition of access mode choice. Attempts on this have been made in this research, but again due to a lack of cases this was not feasible.

However, if additional data would be available, this would be possible. A station could be chosen on the basis of their mode specific qualities. This could increase the insight in which stations are attracting which specific group of passengers. Some stations attract an above average amount of car users who travel the second leg of their journey by train. A station and access mode choice would help explain this behaviour.

REFERENCES

- O'Sullivan, S. & Morral, J., 1996. Walking Distances to and from Light-Rail Transit Stations. *Transportation Research Record: Journal of the Transportation Research Board*, Volume 1538: Pedestrian and Bicycle Research, pp. 19-26.
- Adler, M. W. & van Ommeren, J. M., 2015. *Does Public Transit reduce Car Travel Externalities*, Amsterdam: Tinbergen Institute.
- Akiyama, T. & Okushima, M., 2009. ANALYSIS OF RAILWAY USER TRAVEL BEHAVIOUR PATTERNS OF DIFFERENT AGE GROUPS. *AGE AND MOBILITY*, 23(1), pp. 6-17.
- Babalik-Sutcliffe, E., 2002. Urban rail systems: Analysis of the factors behind success. *Transport Reviews: A Transnational Transdisciplinary Journal*, 22(4), pp. 415-447.
- Batley, R., Dargay, J. & Wardman, M., 2011. The impact of lateness and reliability on passenger rail demand. *Transportation Research Part E: Logistics and Transportation Review*, 47(1), p. 61-72.
- Bent Flyvbjerg, B., Skamris holm, M. K. & Buhl, S. L., 2003. How common and how large are cost overruns in transport infrastructure projects?. *Transport Reviews: A Transnational Transdisciplinary Journal*, Volume 23(1), pp. 71-88.
- Blainey, S., 2010. Trip end models of local rail demand in England and Wales. *Journal of Transport Geography*, Volume 18, p. 153-165.
- Blainey, S. & Evans, S., 2011. *LOCAL STATION CATCHMENTS – RECONCILING THEORY WITH REALITY*. Southampton, s.n.
- Blainey, S. & Mulley, C., 2013. *Using Geographically Weighted Regression to forecast rail demand in the Sydney Region*. Brisbane, Australasian Transport Research Forum.
- Blainey, S. & Preston, J., 2010. *Geographically Weighted Regression Based Analysis of Rail Commuting Around Cardiff*. Lisbon, 12th WCTR.
- Brinckerhoff, P., 1996. Transit and Urban Form. *TCRP Report*, 1(16).
- Brons, M. & Rietveld, P., 2009. Improving the quality of the door-to-door rail journey: a customer-oriented approach.. *Built Environment*, 35(1), pp. 122-135.
- Brown, M., 1983. *Public Transit Fare and Subsidy Policy in Greater Vancouver: Efficiency and Equity Implications*, Vancouver: School of Community and Regional Planning, University of British Columbia.
- Carlier, K., Fiorenzo-Catalano, S., Lindveld, C. & Bovy, P., 2003. *A supernetwork approach towards multimodal travel modeling*. s.l., 82nd Annual Meeting of the Transportation Research Board..
- Carpio-Pinedo, J., 2014. *Urban bus demand forecast at stop level: Space Syntax and other built environment factors. Evidence from Madrid*. Madrid, XI Congreso de Ingenieria del Transporte.
- Cascetta, E. & Cartení, A., 2014. The hedonic value of railways terminals. A quantitative analysis of the impact of stations quality on travellers behaviour.. *Transportation Research*, Volume 61, pp. 41-52.
- Cervero, R., 2006. Alternative approaches to modeling the travel-demand impacts of smart growth.. *Journal of the American Planning Association*, pp. 285-295.
- Cervero, R. & Knockelman, K., 1997. Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, p. 199-219.

- Debrezion, G., Pels, E. & Rietveld, P., 2009. Modelling the joint access mode and railway station choice. *Transportation Research Part E: logistics and transportation review*, 45(1), pp. 270-283.
- Doi, M. & Allen, B. W., 1986. A time series analysis of monthly ridership for an urban rail rapid transit line. *Transportation*, 13(3), pp. 257-269.
- Draak, M., 2010. *De ontbrekende spoorlink van het noorden*, Enschede: Railinfra Solutions, University of Twente.
- Flyvbjerg, B., Skamris Holm, M. K. & Buhl, S. L. B., 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.
- García-Palomares, J. C., Gutiérrez, J. & Cardozo, O. D., 2013. Walking accessibility to public transport: an analysis based on microdata and GIS. *Environment and Planning B: Planning and Design*, 40(6), p. 1087 – 1102 .
- Givoni, M. & Rietveld, P., 2007. *Developing the rail network through better access to railway stations*. Leiden, PROCEEDINGS OF THE EUROPEAN TRANSPORT CONFERENCE 2007 HELD 17-19 OCTOBER 2007.
- Givoni, M. & Rietveld, P., 2014. Do cities deserve more railway stations? The choice of a departure. *Journal of Transport Geography*, Volume 36, pp. 89-97.
- Gutiérrez, J., Cardozo, O. D. & García-Palomares, J. C., 2011. Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. *Journal of Transport Geography*, 19(6), p. 1081–1092.
- Gutiérrez, J. & García-Palomares, J. C., 2008. Distance-measure impacts on the calculation of transport service areas using GIS. *Environment and Planning B: Planning and Design*, 35(3), pp. 480-503..
- Horner, M. W. & Murray, A. T., 2004. Spatial representation and scale impacts in transit service assessment. *Environment and Planning B*, Volume 31, pp. 785-798.
- Keijer, M. & Rietveld, R., 2000. How do people get to the railway station? The Dutch experience. *Transportation Planning and Technology*, 23 (3), p. 215–235.
- Krizek, K. J. & El-Geneidy, A., 2007. Segmenting preferences and habits of transit users and non-users.. *Journal of Public Transportation*, 10.3(71).
- Kuby, M., Barranda, A. & Upch, C., 2004. Factors influencing light-rail station boardings in the. *Transportation Research*, p. 223–247.
- La Paix Puello, L. & Geurs, K., 2015. Modelling observed and unobserved factors in cycling to railway stations: Application to transit-oriented-developments in the Netherlands. *European Journal of Transportation and Infrastructure Research*, 1(15), pp. 27-50.
- Liu, C., Ma, T., Erdogan, S. & Ducca, F. W., 2013. *How to Increase Rail Ridership in Maryland? Direct Ridership Models (DRM) for Policy Guidance*. Maryland, University of Maryland.
- Lythgoe, W. F., Wardman, M. & Toner, J. P., 2004. *Enhancing Rail Passenger Demand Models to examine Station Choice and Access to the Rail Network*. Strasbourg, AET European Transport Conference.

McNally, M. G., 2008. The four step model. In: 2, ed. *Handbook of transport modelling*. Pergamon: Hensher and Buttons.

Ministry of environment and infrastructure, 2014. *Toezegging Nieuwe Stations*, The Hague: Ministry of environment and infrastructure.

Ministry of I&M, 2014. *Kamerbrief over uitkomsten Bestuurlijke Overleggen MIRT 2014: Bijlage 6 Toezegging Nieuwe Stations*, The Hague: Ministry of infrastructure and environment.

Netwerk Zuidelijke Randstad, 2015. *Monitor Regiospoor*, The Hague: Netwerk Zuidelijke Randstad.

NS, Prorail, 2006. *Handleiding PINO*, s.l.: Intern Document.

NS, 2014. *In- en uitstappers per station*. [Online]
Available at: <http://www.Treinreizger.nl>
[Accessed 14 04 2016].

O'Neill, W. A., Douglas, R. R. & JaChing, C., 1992. Analysis of transit service areas using geographic information systems. *Transportation Research Record*, Volume 1364.

Polzin, S., Chu, x. & Rey, J., 2000. Density and captivity in public transit success: observations from the 1995 nationwide personal transportation study. *Transp Res. Rec.: J. Transp. Res. Board*, Volume 1735, p. 10–18.

Preston, J. M., 1987. *The Evaluation of New Local Rail Stations in West Yorkshire*, Leeds: School of Economic Studies (Institute for Transport Studies), University of Leeds.

Preston, J. M. & Dargay, J., 2005. The Dynamics of Rail Demand. *Third Conference on Railroad Industry Structure, Competition and Investment*, pp. 20-22.

Sung, H. & Oh, J.-T. O., 2010. Transit-oriented development in a high-density city: Identifying its association with transit ridership in Seoul, Korea. *Cities*, 28(1), p. 70–82.

Taylor, B. & Fink, C., 2003. *The Factors Influencing Transit Ridership: A Review and Analysis of the Ridership Literature*, Los Angeles: UCLA Department of Urban Planning.

Upchurch, C., Kuby, M., Zoldak, M. & Barranda, A., 2004. Using GIS to generate mutually exclusive service areas linking. *Journal of Transport Geography*, Issue 12, pp. 23-33.

Van Hagen, M. & De Bruyn, M., 2002. *Typisch NS: Elk station zijn eigen rol*. s.l., Colloquium Vervoersplanologisch Speurwerk 2002: De kunst van het verleiden..

Veitch, T. & Cook, J., 2010. *OtTransit: Uses and Functions*, s.l.: Omnitrans International.

Walters, G. & Cervero, R., 2003. *Forecasting Transit Demand in a Fast Growing Corridor: The Direct-Ridership Model Approach. Technical Memorandum prepared for the Bay Area Rapid Transit District*, Lafayette, CA: Fehrs and Peers Associates.

Wardman, M. & Lythgoe, W., 2004. Modelling passenger demand for parkway rail stations. *Transportation*, Volume 31, pp. 125–151,.

Wardman, M., Lythgoe, W. & Whelan, G., 2007. RAIL PASSENGER DEMAND FORECASTING: CROSS-SECTIONAL MODELS REVISITED. *Railroad Economics*, Volume 20, p. 119–152.

Zhao, J., Deng, W., Song, Y. & Zhu, Y., 2013. What influences Metro station ridership in China? Insights from Nanjing. *Cities*, Volume 35, pp. 114-124.

Zhu, X. & Lee, C., 2008. Walkability and safety around elementary schools: economic and ethnic disparities. *American Journal of Preventive Medicine*, Issue 34, p. 282–290.

APPENDICES

APPENDIX 1: PROPOSED STATIONS IN THE NETHERLANDS

Station	Current Appraisal	Estimated Realisation Date
Drenthe		
Assen-Zuid	Not enough demand	After 2028
Flevoland		
Lelystad-Zuid	Not enough demand	After 2028
Friesland		
Leeuwarden-Werpsterhoek	Feasible operation possible	Before 2028
Sneek-Harinxmaland	Possibility for station in future	After 2028
Gelderland		
Apeldoorn West	Not possible within current infrastructure	After 2028
Arnhem's Buiten	Spatially not possible	After 2028
Arnhem Pleij	Not possible within current infrastructure	After 2028
Barneveld Noord	Not possible within current infrastructure	After 2028
Geldermalsen Zuid	Not enough demand	After 2028
Nijkerk Corlaer	Not enough demand	After 2028
Ressen	Not studied yet	After 2028
Stroe	Not enough demand	After 2028
Wijchen West	Not possible within current infrastructure	After 2028
Zevenaar Oost	Not possible within current infrastructure	After 2028
Groningen		
Duurkenakker	Not studied yet	Before 2028
Hoogezand Centrum	Municipality stopped realisation	N.A.
Hoogkerk	Enough demand, depending on other rail project	Before 2028
Sappemeer	Municipality stopped realisation	N.A.
Stadskanaal	Not connected with railway line yet	Before 2028
Wildervank	Not connected with railway line yet	Before 2028
Limburg		
Baaxem	Not enough demand	Before 2028
Belfeld	Not studied yet	After 2028
Haelen	Not enough demand	Before 2028
Maastricht Noord	Study in progress	Before 2028
Venlo Grubbenvorst	Not studied yet	Before 2028
Noord Brabant		
Berkel Enschoot	Not enough demand	After 2028
Breda Oost	Not enough demand	After 2028
Eindhoven Airport	Not enough demand	Before 2028
Oss Oost	Not enough demand	After 2028
Oss West	Not enough demand	After 2028
's-Hertogenbosch Avenue	Not enough demand	After 2028
's-Hertogenbosch Maaspoort	Not enough demand	Before 2028
Overijssel		
De Lutte	Not enough demand	After 2028
Deventer Platvoet	Within current plans unfeasible	After 2028
Deventer Zuid		After 2028
Hengelo Westermaat	Within current plans unfeasible	After 2028
Staphorst	Not enough demand, chances for realisation within line Zwolle-Leeuwarden	Before 2028
Zwolle Stadshagen	Will be constructed in 2017	2017
Zwolle Zuid	Within current plans unfeasible	After 2028
Utrecht		
Amersfoort Koppel	Within current plans unfeasible	After 2028

Amersfoort Oost	Within current plans unfeasible	After 2028
Maartensdijk	Within current plans unfeasible	After 2028
Utrecht Lage Weide	Within current plans unfeasible	Before 2028
Utrecht Majella	Not yet studied	Before 2028
Zuid-Holland		
Boskoop Snijdelwijk	Will be constructed in 2017	2017
Dordrecht Copernicuslaan	Currently in study	Before 2028
Dordrecht Leerpark	Not enough demand	After 2028
Gorinchem Noord	Not enough demand	Before 2028
Hazerswoude Koudekerk	Currently in alternatives study	Before 2028
Leerdam Broekgraaf	Not enough demand	Before 2028
Rotterdam Stadionpark	Not enough demand	After 2028
Schiedam Kethel	Not enough demand	After 2028
Waddinxveen Zuid	Currently in realisation planning	Before 2028
Westergouwe	Not enough demand	After 2028
Zoeterwoude Meerburg	Currently in alternatives study	Before 2028

APPENDIX 2: COMPLETE LIST OF ALL VARIABLES

Variable	Description
Dependent variable:	
Daily_2013	the daily number of passengers boarding or exiting the train at this station
Network quality:	
IC_service	Dummy variable. 1 if full IC service is present, 0 if not
IC_Partial	Dummy variable. 1 if partial IC service is present, 0 if not.
NOL_BTMT	Number of lines for bus, tram or metro with a stop at the station
NOL_BUS	Number of bus lines passing the station
NOL_Stadsbus	Number of city bus lines passing the station
NOL_Streekbus	Number of regional bus lines passing the station
NOL_Metro	Number of metro lines
NOL_tram	Number of tram/light rail lines passing the station
NOL_Bijz	Number of Ferries departing near station
NOL_IC	Number of lines for intercity trains with a stop at the station
NOL_Spr	Number of lines for sprinter trains a stop at the station
Freq_BTMT	Frequency of bus, tram or metro lines with a stop at the station
Freq_Stadsbus	Frequency of city busses
Freq_Streekbus	Frequency of regional busses
Freq_Metro	Frequency of metro
Freq_Tram	Frequency of tram/light rail
Freq_IC	Frequency of lines for intercity trains with a stop at the station
Freq_Spr	Frequency of lines for sprinter trains a stop at the station
Terminal	Dummy indicating if the station is at the end of a line (1) or not (0).
Delay_2013	Number of disruptions in the normal timetable in 2013
Other_Sta	The number of other stations within 15 kilometres of this station
Basis	Accessibility indicator as estimated in Chapter I without any further weighting
SCI	Accessibility indicator as estimated in chapter 1, weighted for the distance ratio rail/road.
CCI	Accessibility indicator as estimated in chapter I weighted for the number of transfers
Af_ONDVMB	Average distance to nearest high school (VMBO)
Af_ONDHV	Average distance to nearest high school (HAVO/VWO)
Af_ONDVRT	Average distance to nearest high school (any)
Af_WARENH	Average distance to nearest department store
Af_Oprith	Average distance to nearest on-ramp to a highway
Af_Overst	Average distance to nearest type 1 or 2 station
Af_BIOS	Average distance to nearest cinema
Af_ATTRACT	Average distance to nearest attraction (museum, amusement park etc.)
Af_Podium	Average distance to nearest theatre
Proximity	Average figure of the combined average distances to a cinema, theatre, department store, type 1 or 2 station and, high school education.
Built environment:	
Randstad	Dummy variable. 1 if station is situated in Randstad area, 0 if not.
Bike_rental	1 if Rental bikes available, 0 otherwise
Bike_park	Dummy. 1 if bicycle parking (self-service or staffed) is available, 0 if not.
Parking_spaces	Number of parking spaces available at the station
PR_Cat	Dummy with Car Parking places: 1 < 50, 2 50-100, 3 100-200, 4 >400
Design	1 = basic station, 2 = station building built before 1945, not in use, 3 = station building built before 1945, still in use, 4 = station built after 1945, 5 = station built after 2000.
Overdekt_perron	Dummy. 1 if one or more platforms are covered, 0 otherwise.
Tot_Potential	Total potential of a station measured in the number of trips including trips from student enrolment, jobs and inhabitants.
Pot_Inw	Potential in the number of trips from inhabitants
Pot_Jobs	Potential in the number of trips from jobs
Pot_Onderwijs	Potential in the number of trips from total student enrolment
Pot_MO	Potential in the number of trips from high school students
Pot_MBO	Potential in the number of trips from lower level higher education
Pot_HBO	Potential in the number of trips from college enrolments
A_Bedv	Total number of businesses
A_Bed_hor_han	Total number of business in the hospitality of small retail sector

A_BED_Fin	Total number of business in the finance sector
A_BED_Zak	Total number of business in the commercial sector
Som_Leisure	Total number of business in the touristic sector
Som_Shop	Total number of retail businesses
OAD	Address density
BEV_DH	Population density
Opp_Bebouwd	Total land area which is developed (not agriculture, pasture etc.)
Detail_Horeca	Total square metres of retail area (shops, restaurants etc.)
Wegverkeersterrein	Total square metres of land used for infrastructure
Woon	Total square metres of residential area
Cultuur	Total square metres of cultural area
Bedrijf	Total square metres of commercial area
Park	Total square metres of (national)park area
Sport	Total square metres of area used or sports
LUM	Land use mix as measured with residential, commercial and retail area
Socio-economic variables:	
Student_Ratio	Ratio of students/total potential in the station area
P_N_W_AL	Percentage of non-western immigrants
WOZ	House value in station area
P_Koopw	Percentage of home owners
P_Leegsw	Percentage of empty/derelict houses
P_WN200	Percentage of houses built after 2000
AUTO_HH	Average number of cars per household
AUTO_LAND	Number of cars per square kilometre
AUTO_BED	Total number of cars used for commercial purposes
AUTO_TOT	Total number of cars registered
Gem_Ink_pi	Average income per inhabitant
P_Ink_Li	Percentage of low income households
P_Ink_Hi	Percentage of high income households
P_0014	Percentage of people aged between 0-14
P_15-34	Percentage of people aged between 15-34
P_35-64	Percentage of people aged between 35-64
P_65-75	Percentage of people aged between 65-75
P_75oud	Percentage of people aged over 75 years old
P1P_HH	Percentage of household consisting of only 1 person
MP_HH_ZK	Percentage of households consisting of multiple persons and no children
MP_HH_MK	Percentage of households consisting of multiple persons and with children
HH_GRT	Average household size
P_NIETACT	Percentage of non-active persons (retired, unemployed etc.)

APPENDIX 3: OVERVIEW OF MNL STATION CHOICE MODEL 1

Name	Value	Robust Std err	Robust t-test	p-value
ASC_1	0.00			
ASC_2	-2.12	0.441	-4.81	0.00
ASC_3	-4.03	0.825	-4.89	0.00
ASC_4	-4.60	0.759	-6.06	0.00
ASC_5	-4.78	1.01	-4.75	0.00
ASC_other	-1.94	1.35	-1.44	0.15
BTM1	0.0284	0.00578	4.91	0.00
BTM2	0.0277	0.00701	3.95	0.00
BTM3	0.0112	0.0135	0.83	0.40
BTM4	0.0230	0.0126	1.82	0.07
BTM5	0.00			
BTM6	0.0117	0.0235	0.50	0.62
Bike1	0.0467	0.0132	3.53	0.00
Bike2	0.00			
Bike3	0.00			
Bike4	0.00			
Bike5	0.00			
Bike6	0.213	0.0840	2.53	0.01
Dist1	0.00			
Dist2	-0.00143	0.000152	-9.36	0.00
Dist3	-0.000789	0.000217	-3.64	0.00
Dist4	-0.000466	0.000145	-3.20	0.00
Dist5	-0.000197	0.000119	-1.65	0.10
Dist6	-0.000393	0.000163	-2.41	0.02
Distance1	0.00			
Distance2	0.00165	0.000162	10.17	0.00
Distance3	0.00109	0.000211	5.19	0.00
Distance4	0.000933	0.000162	5.77	0.00
Distance5	0.000591	0.000170	3.48	0.00
Distance6	0.000842	0.000110	7.66	0.00
Frequency1	0.00			
Frequency2	0.0396	0.00893	4.44	0.00
Frequency3	0.0751	0.0193	3.89	0.00
Frequency4	0.0531	0.0157	3.37	0.00
Frequency5	0.0522	0.00982	5.31	0.00
Frequency6	-0.106	0.0372	-2.83	0.00
RAIL_acces1	0.00			
RAIL_acces2	1.19	0.376	3.16	0.00
RAIL_acces3	1.56	0.391	4.00	0.00
RAIL_acces4	1.57	0.416	3.78	0.00
RAIL_acces5	1.16	0.559	2.07	0.04
RAIL_acces6	1.18	0.590	2.00	0.05
other_St	0.0661	0.0275	2.40	0.02

APPENDIX 4: POTENTIAL FOR SPRINTER STATIONS

Station Name	Actual Ridership	Tot_Pot	Pot_Inw	Pot_Jobs	Pot_Ond	Pot_MO	Pot_MBO	Pot_HBO
Aalten	1341	2031	1462	227	0	0	0	0
Abcoude	1625	1081	708	354	19	0	0	19
Akkrum	719	565	486	80	0	0	0	0
Alkmaar Noord	4950	7918	3065	3240	868	0	582	286
Almelo de Riet	1242	1548	832	548	109	0	109	0
Almere Buiten	7900	4612	2895	802	578	41	537	0
Almere Muziekwijk	7030	3270	1810	1091	289	5	232	52
Almere Oostvaarders	4285	1954	1278	385	21	2	19	0
Almere Parkwijk	3907	2206	1478	440	89	5	0	84
Almere Poort	2256	1383	620	449	38	0	37	0
Amersfoort Schothorst	5642	3714	1759	1432	188	3	173	11
Amersfoort Vathorst	2559	2214	1362	599	12	11	0	0
Amsterdam Bijlmer ArenA	18961	10860	4514	5703	644	184	310	150
Amsterdam Holendrecht	3176	4324	1615	2065	612	0	0	612
Amsterdam Lelylaan	12469	10635	6650	3103	404	4	175	225
Amsterdam Muiderpoort	11147	9951	5850	3213	537	90	21	425
Amsterdam RAI	6273	5835	2773	2283	657	137	468	52
Amsterdam Sciencepark	3225	872	388	390	27	2	25	0
Anna Paulowna	2333	861	775	86	0	0	0	0
Apeldoorn De Maten	619	1473	934	403	20	4	9	7
Apeldoorn Osseveld	1040	1715	1170	506	21	12	5	5
Appingedam	1106	1686	1066	296	55	0	55	0
Arkel	402	581	429	147	5	5	0	0
Arnhemuiden	488	870	603	267	0	0	0	0
Arnhem Presikhaaf	3162	2734	802	825	1043	34	164	845
Arnhem Velperpoort	3672	2021	925	700	251	14	135	102
Arnhem Zuid	2790	2079	1595	460	1	0	0	0
Baarn	4658	2914	1960	836	55	55	0	0
Bafo	666	301	259	42	0	0	0	0
Barendrecht	4973	4160	2492	1289	22	0	22	0
Barneveld Centrum	3010	3115	1791	888	157	142	13	2
Barneveld Noord	1231	827	180	506	142	40	0	102
Barneveld Zuid	900	1084	493	327	12	0	12	0
Bedum	483	876	716	160	0	0	0	0
Beek-Elst	2258	2170	1376	686	7	0	7	0
Beesd	183	520	272	248	0	0	0	0
Beilen	2064	1133	902	109	0	0	0	0
Bergen op Zoom	7220	8266	4240	2228	858	0	505	353
Best	5322	3997	2442	1247	0	0	0	0
Bilthoven	4380	3171	1717	950	0	0	0	0
Blerick	1101	2668	1399	814	243	0	230	13
Bloemendaal	1385	1201	668	201	91	14	72	5
Bodegraven	3005	2547	1891	656	0	0	0	0
Borne	2348	2012	1601	345	0	0	0	0
Boskoop	1428	1305	915	318	0	0	0	0
BovenHardinxveld	343	546	433	113	0	0	0	0
Bovenkarspel Flora	867	779	509	232	39	39	0	0
Bovenkarspel-Grootebroek	2399	1846	1283	308	0	0	0	0
Boxmeer	4093	3796	1317	885	665	0	665	0
Boxtel	6325	3561	2125	798	309	0	309	0
Breda Prinsenbeek	1260	2363	1517	763	4	0	1	2
Bruckelen	5058	1226	750	248	0	0	0	0
Brummen	1075	1005	812	193	0	0	0	0
Buitenpost	1941	1215	731	134	65	0	65	0
Bunde	954	752	573	173	6	6	0	0
Bunnik	2005	719	441	261	17	0	0	17
Bussum Zuid	3907	1086	535	508	7	7	0	0
Capelle Schollevaar	2242	1913	1610	198	47	23	25	0
Chevremont	586	1030	784	232	14	14	0	0
Coevorden	1866	1849	1286	346	0	0	0	0
Cuijk	3497	2692	1680	821	0	0	0	0
Culemborg	8232	3886	2293	844	0	0	0	0
Daarlerveen	116	344	249	64	30	5	26	0
Dalen	202	450	401	48	1	1	0	0
Dalfsen	1533	754	633	121	0	0	0	0
De Vink	2783	1533	1233	162	94	17	58	19
Deinum	137	233	151	81	1	0	1	0
Delden	904	1133	675	403	0	0	0	0
Delft Zuid	4668	2058	875	659	494	0	0	493
Delfzijl	1162	1042	668	301	74	0	74	0
Delfzijl West	442	1118	684	408	26	16	10	0
Den Dolder	1942	846	444	402	0	0	0	0
Den Haag Mariahoeve	2877	1731	1056	553	46	8	37	1
Den Haag Moerwijk	2296	3084	2856	167	25	4	18	2
Den Haag Ypenburg	1801	1835	1369	285	1	1	0	0
Den Helder Zuid	1918	1216	651	403	64	0	64	0
Deurne	4703	2408	1852	284	14	0	14	0
Deventer Colmschate	1646	2478	1766	658	11	0	4	7
Didam	1899	2120	1237	595	0	0	0	0
Diemen	3423	2180	1485	672	22	0	22	0
Diemen Zuid	3304	1726	739	721	255	2	142	111
Doetinchem	3968	4391	1494	1239	1088	124	912	52
Doetinchem De Huet	1213	1706	763	608	230	76	146	8
Dordrecht Stadspolders	709	1785	1326	404	13	1	12	0
Dordrecht Zuid	1241	1601	953	444	75	6	68	1
Driebergen-Zeist	9267	4248	2487	1305	3	0	3	0
Driehuis	974	1667	1155	210	0	0	0	0
Dronrijp	155	299	223	76	0	0	0	0
Dronten	3142	1975	1476	119	92	0	24	68
Duiven	3865	3291	2199	626	0	0	0	0
Echt	2356	1832	1412	263	0	0	0	0

Ede Centrum	1084	2239	1432	689	51	0	39	13
Eijsden	213	1006	810	197	0	0	0	0
Eindhoven Beukenlaan	1938	3179	1241	1482	377	0	240	137
Elst	3863	1940	1360	335	122	0	122	0
Emmen	2436	4978	2029	1952	251	23	19	209
Emmen Zuid	698	1243	820	409	1	0	1	0
Enkhuizen	2604	1770	1244	402	0	0	0	0
Enschede De Eschmarke	81	589	470	107	0	0	0	0
Enschede Drienerlo	2976	1692	242	445	974	1	267	706
Ermelo	2904	2998	1880	785	0	0	0	0
Eygelshoven	312	571	306	254	12	12	0	0
Eygelshoven Markt	285	773	518	255	0	0	0	0
Franeker	918	1441	1064	256	0	0	0	0
Gaanderen	339	1008	653	348	8	8	0	0
Geerdijk	93	353	231	122	0	0	0	0
Geldermalsen (NS+Arriva)	5856	1752	1218	312	68	0	68	0
Geldrop	1555	3004	1976	931	0	0	0	0
Geleen Oost	600	1321	848	380	17	1	16	0
Geleen-Lutterade	1504	1435	809	568	53	24	29	0
Gilze-Rijen	2616	2235	1800	434	0	0	0	0
Glanerbrug	308	1050	949	100	1	1	0	0
Goes	7660	7084	2579	2844	681	0	681	0
Goor	1599	1956	1229	670	0	0	0	0
Gorinchem	4113	5246	2704	1707	24	0	24	0
Gouda Goverwelle	2835	1119	807	286	2	0	2	0
Gramsbergen	289	448	398	50	0	0	0	0
Grijpskerk	840	478	352	80	0	0	0	0
Groningen Europapark	989	981	341	453	139	15	78	47
Groningen Noord	1701	3415	1615	669	914	1	147	765
Grou-Jirnsom	923	666	560	63	0	0	0	0
Haarlem Spaarnwoude	3086	1683	661	978	35	2	33	0
Halfweg	1478	1200	808	392	0	0	0	0
Harde 't	1236	905	593	312	0	0	0	0
Hardenberg	3175	3502	1518	179	1078	0	121	957
Harderwijk	5992	6125	3338	1435	306	0	306	0
Hardinxveld Blauwe Zoom	246	788	438	201	0	0	0	0
Hardinxveld-Giessendam	660	846	753	82	11	11	0	0
Haren	1132	1338	970	152	0	0	0	0
Harlingen	1840	1594	1007	235	118	0	118	0
Harlingen Haven	341	553	387	103	63	44	18	0
Heemskerk	2267	2345	1836	315	0	0	0	0
Heerhugowaard	7818	5879	3389	1089	814	0	814	0
Heerlen de Kissel	371	724	367	226	101	3	97	2
Heerlen Woonboulevard	85	272	45	197	26	2	18	6
Heeze	1634	1221	1080	141	0	0	0	0
Heiloo	4614	2670	2040	596	0	0	0	0
Heino	710	626	473	153	0	0	0	0
Helmond Brandevoort	1021	1694	778	699	4	0	4	0
Helmond Brouwhuis	2057	1954	1171	537	117	2	0	115
Helmond 't Hout	1247	1566	948	568	50	17	33	0
Hemmen-Dodewaard	141	212	155	56	0	0	0	0
Hengelo Gezondheidspark	1450	1009	410	528	22	3	15	4
Hengelo Oost	2500	1193	748	323	75	1	59	15
Hertogenbosch 's Oost	1764	2553	863	1226	148	0	4	144
Hillegom	2429	2253	1446	692	0	0	0	0
Hilversum Noord	3795	786	269	465	8	5	3	0
Hilversum Sportpark	7208	5439	2565	1677	568	43	415	110
Hindeloopen	115	106	65	41	0	0	0	0
Hoek van Holland Haven	1569	724	678	45	1	1	0	0
Hoek van Holland Strand	494	360	289	61	0	0	0	0
Hoensbroek	196	617	276	323	0	0	0	0
Hoevelaken	1500	560	286	216	0	0	0	0
Hollandsche Rading	852	755	258	490	0	0	0	0
Holten	1290	1420	757	186	0	0	0	0
Hoofddorp	15068	7183	4076	2030	821	0	232	589
Hoogeveen	4328	4490	2050	1785	332	0	332	0
Hoogezand-Sappemeer	1272	1522	1021	369	6	6	0	0
Hoogkarspel	2300	1114	900	212	2	2	0	0
Hoorn Kersenboogerd	5388	2697	1677	904	9	9	1	0
Horst-Sevenum	2635	1305	941	232	41	0	41	0
Houten	7478	4382	2356	1276	80	0	80	0
Houten Castellum	3499	3119	2050	1044	25	25	0	0
Houthem-St. Gerlach	341	438	295	140	3	3	0	0
Hurdegaryp	1001	710	600	98	0	0	0	0
IJlst	260	430	379	51	0	0	0	0
Kampen	4256	3741	2952	531	73	68	6	0
Kampen Zuid	1141	743	348	146	152	3	150	0
Kapelle-Biezelinge	995	1301	920	168	101	0	0	101
Kerkrade Centrum	1115	1409	1058	319	0	0	0	0
Kesteren	505	1013	532	260	0	0	0	0
Klarenbeek	283	224	137	88	0	0	0	0
Klimmen-Ransdaal	373	269	212	57	0	0	0	0
Koog Bloemwijk	3016	1171	756	319	79	57	15	7
Koog-Zaandijk	3072	775	457	202	0	0	0	0
Koudum-Molkwerum	160	175	132	37	0	0	0	0
Krabbendijke	578	829	553	96	0	0	0	0
Krommenie-Assendelft	5640	2467	1551	402	0	0	0	0
Kropswolde	529	577	297	281	0	0	0	0
Kruiningen-Yerseke	874	474	353	121	0	0	0	0
Lage Zwaluwe	767	348	157	191	0	0	0	0
Landgraaf	1228	795	489	178	1	0	1	0
Leerdam	712	3173	2096	850	0	0	0	0
Leeuwarden Camminghaburen	835	1838	643	878	275	0	269	5
Leiden Lammenschans	3643	3327	1276	747	869	9	517	343

Lichtenvoorde-Groenlo	878	1133	710	390	0	0	0	0
Lochem	1286	1059	854	152	0	0	0	0
Loppersum	602	432	394	38	0	0	0	0
Lunteren	1044	1080	987	93	0	0	0	0
Maarheeze	1258	599	418	181	0	0	0	0
Maarn	1507	1129	847	276	0	0	0	0
Maarsse	4744	4127	2238	1541	1	1	0	0
Maassluis	2099	1321	1141	89	21	8	13	0
Maassluis West	2510	1585	1174	286	97	12	85	0
Maastricht Randwyck	3672	1959	699	883	261	2	48	210
Mantgum	495	258	212	46	0	0	0	0
Marienberg	333	380	234	145	0	0	0	0
Martenshoek	1013	1449	945	408	29	29	0	0
Meerssen	1223	954	521	156	0	0	0	0
Meppel	5346	3930	2276	1214	51	0	51	0
Middelburg	4800	4591	2913	1059	95	0	95	0
Mook Molenhoek	1224	1196	963	186	0	0	0	0
Naarden-Bussum	9778	5690	3391	1584	127	127	0	0
Nieuw Amsterdam	784	830	640	190	0	0	0	0
Nieuw Vennep	2556	2134	1639	347	0	0	0	0
Nieuwerkerk a/d IJssel	3306	2235	1761	390	3	0	3	0
Nieuweschem	588	201	183	17	0	0	0	0
Nijkerk	3650	2601	1962	437	63	0	63	0
Nijmegen Dukenburg	2151	2195	1209	919	5	5	0	0
Nijmegen Goffert	1000	1541	648	734	62	5	5	51
Nijmegen Heyendaal	3287	6141	1064	1930	2815	30	13	2772
Nijmegen Lent	925	645	378	124	2	0	0	1
Nijverdal	2939	2913	2050	614	0	0	0	0
Nunspeet	2824	2245	1531	664	0	0	0	0
Nuth	453	947	545	193	209	6	203	0
Obdam	1558	1020	828	192	0	0	0	0
Oisterwijk	2369	2658	1879	487	0	0	0	0
Oldenzaal	3275	4375	2258	1561	0	0	0	0
Olst	1293	858	781	77	0	0	0	0
Ommen	1910	858	657	93	0	0	0	0
Oosterbeek	482	830	537	288	4	3	1	0
Opheusden	391	775	539	236	0	0	0	0
Oss West	1806	1169	926	173	52	14	38	0
Oudenbosch	1257	2502	1476	593	0	0	0	0
Overveen	3092	1052	389	245	381	13	133	235
Purmerend	2992	4216	2470	1481	265	211	53	0
Purmerend Overwhere	2312	4653	2068	1229	944	0	944	0
Purmerend Weidevenne	1646	2325	1813	512	0	0	0	0
Putten	1914	1360	1127	213	0	0	0	0
Raalte	2059	2847	1585	364	342	0	342	0
Ravenstein	1364	675	564	82	0	0	0	0
Reuver	1519	1537	1240	226	0	0	0	0
Rheden	907	1024	852	171	0	0	0	0
Rhenen	1401	1333	1062	216	19	19	0	0
Rijssen	2428	3710	2274	852	349	0	349	0
Rijswijk	7141	7930	4136	3269	52	9	43	0
Rilland-Bath	437	377	237	140	1	1	0	0
Roodeschool	247	190	161	30	0	0	0	0
Rosmalen	2324	3391	2162	924	1	1	0	0
Rotterdam Lombardijen	6272	4263	2571	621	548	55	492	0
Rotterdam Noord	2302	1017	694	191	63	5	46	12
Rotterdam Zuid	2799	1819	1436	166	134	14	115	5
Ruurlo	951	787	677	110	0	0	0	0
Santpoort Noord	864	816	667	123	26	26	0	0
Santpoort Zuid	866	650	500	113	23	17	6	0
Sappemeer Oost	551	733	412	321	0	0	0	0
Sassenheim	3000	2241	1308	692	0	0	0	0
Sauwerd	372	222	184	38	0	0	0	0
Schagen	5921	3289	1855	419	220	0	220	0
Scheemda	694	558	510	47	1	1	0	0
Schiedam Nieuwland	4835	1270	757	222	96	3	93	0
Schin op Geul	371	240	163	77	0	0	0	0
Schinnen	346	455	331	119	4	0	4	0
Slidrecht	866	579	276	294	1	1	0	0
Slidrecht Baanhoek	553	1798	752	1018	4	4	0	0
Sneek	2901	3956	1893	778	191	0	191	0
Sneek Noord	959	1103	713	324	66	44	22	0
Soest	231	403	281	122	0	0	0	0
Soest Zuid	2060	1214	1044	169	0	0	0	0
Soestdijk	938	1358	784	484	1	1	0	0
Spaubeek	397	505	288	216	1	1	0	0
Stavoren	331	149	126	22	0	0	0	0
Stedum	287	154	122	32	0	0	0	0
Susteren	928	908	748	160	0	0	0	0
Swalmen	453	893	688	174	0	0	0	0
Tegelen	745	1946	1325	610	11	2	1	8
Terborg	711	1184	838	213	0	0	0	0
Tiel	4128	5490	2591	2199	227	0	227	0
Tiel Passewaaij	1269	1110	880	225	5	4	2	0
Tilburg Reeshof	2563	3699	2130	885	3	0	0	3
Tilburg Universiteit	7348	6110	1716	1387	2742	0	337	2405
Twello	1554	1564	1159	163	93	0	93	0
Uitgeest	5336	2205	1813	390	2	2	0	0
Uithuizen	891	709	591	43	0	0	0	0
Uithuizermeeden	420	334	301	33	0	0	0	0
Usquert	226	229	182	47	0	0	0	0
Utrecht Leidsche Rijn	4700	222	129	79	0	0	0	0
Utrecht Lunetten	3458	388	195	133	11	0	1	9
Utrecht Overvecht	6827	3418	1625	685	951	0	121	829

Utrecht Terwijde	2626	772	561	199	12	12	0	0
Utrecht Zuilen	1918	390	206	148	35	1	16	18
Valkenburg	1691	1071	787	210	0	0	0	0
Varsseveld	533	976	675	300	0	0	0	0
Veendam	2000	2480	1652	457	86	0	86	0
Veenendaal Centrum	2438	5244	2871	1627	163	116	47	0
Veenendaal West	1549	3005	1526	953	99	34	65	0
Veenendaal-de Klomp	3745	1444	787	656	0	0	0	0
Veenwouden	982	658	552	106	0	0	0	0
Velp	1625	1893	1019	532	203	0	133	70
Venray	3295	2786	1679	745	215	0	215	0
Vierlingsbeek	596	498	391	107	0	0	0	0
Vlaardingen Centrum	2790	1229	991	205	33	24	7	1
Vlaardingen Oost	2817	3952	2292	1057	238	98	24	116
Vlaardingen West	1951	1559	764	451	80	0	80	0
Vleuten	3334	1578	1175	318	0	0	0	0
Viissingen Souburg	1007	1594	1174	299	39	3	12	24
Voerendaal	346	311	232	80	0	0	0	0
Voorburg	2050	3442	1807	1278	25	22	0	3
Voorhout	3452	2513	1819	603	2	2	0	0
Voorschoten	2917	1323	1052	264	7	7	0	0
Voorst-Empe	342	350	229	122	0	0	0	0
Vorden	1051	811	635	109	0	0	0	0
Vriezenveen	311	990	813	133	0	0	0	0
Vroomshoop	341	990	660	107	117	0	117	0
Vught	2010	2523	1647	616	2	1	1	0
Waddinxveen	1575	1882	1225	443	83	0	83	0
Waddinxveen Noord	1238	1237	918	293	26	19	7	0
Warffum	695	420	229	46	0	0	0	0
Weesp	9440	2833	1952	676	0	0	0	0
Wehl	567	944	725	220	0	0	0	0
Westervoort	2250	1314	826	464	25	4	0	20
Wezep	1067	1132	923	186	0	0	0	0
Wierden	1837	1258	1051	172	0	0	0	0
Wijchen	4214	4009	2759	947	0	0	0	0
Wijhe	1125	747	595	81	0	0	0	0
Winschoten	2447	3340	1645	1212	86	0	86	0
Winsum	2382	1083	813	52	0	0	0	0
Winterswijk	1935	2351	1381	645	11	11	0	0
Winterswijk West	372	1266	532	646	65	65	0	0
Woerden	11648	5442	2927	1217	278	0	278	0
Wolfheze	542	575	295	257	0	0	0	0
Wolvega	1521	1645	1173	195	0	0	0	0
Workum	476	407	352	56	0	0	0	0
Wormerveer	4091	2815	1578	1203	34	34	0	0
Zaandam Kogerveld	1713	1074	658	250	19	13	4	2
Zaltbommel	3417	2105	1007	929	0	0	0	0
Zandvoort aan Zee	5200	1923	1635	267	0	0	0	0
Zetten-Andelst	732	1084	695	212	0	0	0	0
Zevenaar	4652	2265	1812	279	4	4	0	0
Zevenbergen	1263	1915	1235	437	0	0	0	0
Zoetermeer	5947	7143	4542	2020	463	261	58	144
Zoetermeer Oost	3196	881	423	328	18	5	12	1
Zuidbroek	809	494	440	54	0	0	0	0
Zuidhorn	2595	1063	931	86	0	0	0	0
Zwaagwesteinde	614	825	728	97	0	0	0	0
Zwijndrecht	5264	5926	3590	1970	41	6	0	35

APPENDIX 5: INTER-CORROLATION BETWEEN VARIABLES

	Tot_Potentie	Pot_Inw	Pot_Jobs	Pot_Onderwijs	Pot_MO	Pot_MBO	Pot_HBO	Parking	PR_CAT	Parking_spaces	Bicycle_parking	Bicycle_rental	Delay_2013	IC_Freq	Sprinter_Freq	Freq_Tot	Freq_BT	Stadsbus_Freq	Streekbus_Freq	Tram_Freq	Metro_Freq	IC_NOL	Sprinter_NOL	Stadsbus_NOL	Streekbus_NOL	Tram_NOL	Metro_NOL	BTM_NOL	Stadsvervoer_Freq	CCI_2013_Actual	P_woon	Opp_bebouwd	Bev_DH	P_N_W_AL	P_KOOPWON	AUTO_TOT	AUTO_HH	AUTO_LAND	BEDR_AUTO	P_3464	PIP_HH	MP_HH_ZK	HH_GRT	A_PART_HH	A_BEDV	A_BED_Hor_Handel	Overdekt_perron	Design_modern	AV3_ONDHW	AV3_ONDHW	AV3_ONDHW	AV3_ONDVB	AV3_ONDVB	AV3_ONDVB	AV3_ONDVRT	AV3_ONDVRT	AV3_WARENH	AV3_WARENH	AV3_WARENH	AV3_WARENH	Proximity	OAD
Tot_Potentie	1.00	0.94	0.92	0.54	0.40	0.53	0.33	0.06	0.45	0.44	0.63	0.39	0.32	0.25	0.29	0.44	0.73	0.44	0.58	0.37	0.33	0.24	0.21	0.43	0.53	0.30	0.34	0.66	0.49	0.42	0.47	0.58	0.55	0.44	-0.39	0.50	-0.42	0.52	0.40	0.24	0.38	-0.32	0.38	0.43	0.47	0.51	0.40	0.08	0.39	0.42	-0.40	0.40	0.53	-0.42	0.50	-0.42	-0.35	0.35	-0.37	-0.37	0.31	0.57
Pot_Inw	0.94	1.00	0.79	0.31	0.38	0.38	0.13	0.08	0.47	0.44	0.61	0.40	0.34	0.24	0.28	0.42	0.67	0.45	0.50	0.42	0.28	0.23	0.19	0.44	0.46	0.35	0.27	0.58	0.49	0.44	0.51	0.56	0.57	0.41	-0.31	0.49	-0.32	0.51	0.37	0.21	0.26	-0.26	0.25	0.44	0.45	0.47	0.42	0.07	0.31	0.32	-0.37	0.32	0.43	-0.39	0.40	-0.39	-0.37	0.31	-0.33	-0.31	0.55	
Pot_Jobs	0.92	0.79	1.00	0.46	0.44	0.44	0.27	0.01	0.36	0.39	0.61	0.35	0.28	0.24	0.26	0.40	0.70	0.48	0.50	0.31	0.46	0.22	0.20	0.46	0.44	0.24	0.47	0.58	0.53	0.43	0.36	0.48	0.46	0.43	-0.42	0.44	-0.42	0.48	0.37	0.26	0.40	-0.35	0.46	0.40	0.24	0.28	0.29	0.08	0.00	0.39	0.42	-0.21	0.37	0.44	-0.22	0.45	-0.22	-0.14	0.25	-0.25	-0.25	0.37
Pot_Onderwijs	0.54	0.31	0.46	1.00	0.23	0.61	0.88	-0.02	0.18	0.18	0.23	0.09	0.09	0.10	0.20	0.28	0.42	0.14	0.42	0.11	0.18	0.10	0.16	0.13	0.31	0.11	0.18	0.34	0.18	0.13	0.19	0.34	0.30	0.28	-0.35	0.29	-0.43	0.28	0.22	0.16	0.44	-0.29	0.40	0.24	0.28	0.29	0.08	0.00	0.39	0.42	-0.21	0.37	0.44	-0.22	0.45	-0.22	-0.14	0.25	-0.25	-0.25	0.37	
Pot_MO	0.40	0.38	0.44	0.23	1.00	0.18	0.06	0.03	0.14	0.20	0.31	0.10	0.17	0.00	0.20	0.22	0.44	0.40	0.29	0.13	0.30	-0.03	0.15	0.38	0.17	0.15	0.31	0.28	0.37	0.22	0.17	0.22	0.29	-0.19	0.30	-0.20	0.28	0.20	0.08	0.17	-0.18	0.22	0.21	0.32	0.31	0.19	0.02	0.23	0.30	-0.16	0.23	0.28	-0.16	0.29	-0.16	-0.12	0.19	-0.16	-0.19	0.32		
Pot_MBO	0.53	0.38	0.44	0.61	0.18	1.00	0.16	0.03	0.22	0.28	0.32	0.19	0.09	0.22	0.09	0.21	0.37	0.12	0.37	0.11	0.16	0.23	0.07	0.12	0.40	0.12	0.16	0.43	0.15	0.10	0.19	0.34	0.24	0.24	-0.26	0.25	-0.31	0.24	0.26	0.14	0.27	-0.17	0.21	0.22	0.31	0.10	-0.02	0.29	0.36	-0.22	0.29	0.41	-0.25	0.40	-0.25	-0.16	0.20	-0.23	-0.22	0.30		
Pot_HBO	0.33	0.13	0.27	0.88	0.06	0.16	1.00	-0.04	0.09	0.04	0.06	-0.01	0.05	-0.01	0.17	0.18	0.26	0.06	0.28	0.06	0.10	-0.01	0.15	0.05	0.13	0.04	0.10	0.15	0.09	0.08	0.11	0.20	0.21	0.19	-0.27	0.19	-0.34	0.16	0.10	0.12	0.37	-0.24	0.36	0.18	0.18	0.15	0.02	0.01	0.30	0.28	-0.11	0.27	0.28	-0.12	0.30	-0.12	-0.07	0.17	-0.17	-0.17	0.26	
Parking	0.06	0.08	0.01	-0.02	0.03	0.03	-0.04	1.00	0.26	0.25	0.04	0.07	0.00	0.09	-0.07	-0.03	-0.01	-0.10	0.08	-0.07	-0.06	0.07	-0.11	-0.08	0.14	-0.08	-0.06	0.11	-0.11	-0.04	-0.01	0.00	-0.15	-0.21	0.13	-0.16	0.16	-0.07	-0.18	-0.25	-0.14	0.21	-0.22	-0.28	-0.26	-0.23	-0.02	-0.10	-0.20	-0.08	0.05	-0.18	-0.09	0.06	-0.11	0.07	0.04	-0.24	0.14	0.18	-0.17	
PR_CAT	0.45	0.47	0.36	0.18	0.14	0.22	0.09	0.26	1.00	0.89	0.37	0.42	0.31	0.28	0.18	0.34	0.33	0.14	0.35	0.05	0.11	0.28	0.07	0.15	0.32	0.05	0.11	0.35	0.13	0.44	0.21	0.25	0.24	0.18	-0.09	0.27	-0.14	0.27	0.19	0.09	0.09	-0.17	0.12	0.09	0.17	0.21	0.32	0.10	0.15	0.20	-0.23	0.19	0.25	-0.25	0.25	-0.25	-0.28	0.13	-0.17	-0.17	0.22	
Parking_spaces	0.44	0.44	0.39	0.18	0.20	0.28	0.04	0.25	0.89	1.00	0.43	0.36	0.27	0.30	0.16	0.33	0.36	0.17	0.37	0.01	0.15	0.29	0.07	0.18	0.32	0.01	0.15	0.35	0.14	0.42	0.13	0.20	0.17	-0.09	0.23	-0.14	0.21	0.20	0.06	0.10	-0.15	0.14	0.08	0.15	0.19	0.27	0.09	0.12	0.17	0.19	0.14	0.20	-0.20	0.19	-0.20	-0.22	0.09	-0.14	-0.17	0.17		
Bicycle_parking	0.63	0.61	0.61	0.23	0.31	0.32	0.06	0.04	0.37	0.43	1.00	0.40	0.27	0.15	0.26	0.36	0.45	0.30	0.40	0.10	0.15	0.17	0.18	0.31	0.38	0.09	0.16	0.44	0.26	0.36	0.19	0.24	0.27	0.19	-0.18	0.25	-0.20	0.29	0.24	0.13	0.23	0.20	-0.15	0.16	0.29	0.28	0.46	0.15	0.20	0.25	-0.20	0.19	0.29	-0.21	0.30	-0.21	-0.16	0.20	-0.10	-0.19	0.32	
Bicycle_rental	0.39	0.40	0.35	0.09	0.10	0.19	-0.01	0.07	0.42	0.36	0.40	1.00	0.29	0.25	0.17	0.31	0.22	0.18	0.16	0.11	0.20	0.24	0.06	0.19	0.18	0.09	0.09	0.22	0.17	0.51	0.23	0.23	0.29	0.16	-0.16	0.24	-0.24	0.34	0.21	0.14	0.19	-0.18	0.18	0.15	0.27	0.24	0.34	0.13	0.23	0.18	-0.19	0.25	0.25	-0.18	0.26	-0.18	-0.23	0.22	-0.26	-0.25	0.29	
Delay_2013	0.25	0.24	0.28	0.09	0.17	0.09	0.05	0.00	0.31	0.27	0.27	0.29	1.00	-0.02	0.54	0.56	0.34	0.32	0.16	0.31	0.19	-0.04	0.36	0.34	0.05	0.30	0.20	0.16	0.36	0.63	0.19	0.23	0.40	-0.19	0.31	-0.26	0.39	0.24	0.38	0.34	0.28	-0.41	0.32	0.43	0.40	0.35	0.37	0.14	0.36	0.29	0.28	0.33	0.27	-0.18	0.25	-0.17	-0.18	0.31	-0.24	-0.30	0.37	
IC_Freq	0.32	0.24	0.24	0.10	0.10	0.00	0.22	-0.01	0.09	0.28	0.30	0.15	0.25	-0.02	1.00	-0.36	0.14	0.17	0.06	0.15	0.06	0.14	0.96	-0.31	0.05	0.25	0.01	0.24	0.26	0.10	0.07	0.02	0.04	0.01	0.02	0.01	-0.01	-0.06	-0.02	0.01	0.04	0.03	0.06	0.05	0.06	-0.02	0.01	0.04	0.01	0.04	0.10	-0.05	0.06	-0.05	0.02	-0.01	0.00	0.03	0.02			
Sprinter_Freq	0.29	0.28	0.26	0.20	0.20	0.09	0.17	-0.07	0.18	0.16	0.26	0.17	0.54	-0.36	1.00	0.87	0.34	0.27	0.24	0.21	0.10	-0.35	0.88	0.28	0.01	0.22	0.10	0.10	0.26	0.53	0.23	0.27	0.43	-0.48	-0.30	0.38	-0.37	0.46	0.28	0.34	0.28	-0.41	0.32	0.43	0.40	0.35	0.37	0.14	0.36	0.29	0.28	0.33	0.27	-0.22	0.29	-0.21	-0.22	0.32	-0.33	-0.33	0.42	
Freq_Tot	0.44	0.42	0.40	0.26	0.22	0.21	0.18	-0.03	0.07	0.28	0.29	0.17	0.24	-0.04	0.96	-0.35	0.13	0.46	0.31	0.34	0.25	0.18	0.13	0.77	0.32	0.14	0.24	0.18	0.24	0.33	0.60	0.26	0.31	0.46	0.52	0.31	0.39	-0.42	0.47	0.38	0.32	-0.45	0.39	0.47	0.43	0.40	0.42	0.14	0.39	0.33	0.29	0.37	0.36	-0.25	0.34	-0.25	0.33	-0.35	-0.34	0.46		
Freq_BT	0.73	0.67	0.70	0.42	0.44	0.37	0.26	-0.01	0.33	0.36	0.45	0.22	0.34	0.17	0.73	0.34	0.51	0.60	0.67	0.77	0.48	0.42	0.14	0.28	0.65	0.60	0.44	0.47	0.79	0.70	0.44	0.24	0.33	0.45	-0.58	-0.44	0.43	-0.45	0.39	0.29	0.29	-0.42	0.45	0.52	0.44	0.45	0.40	0.07	0.42	0.39	-0.26	0.44	0.49	-0.27	0.45	-0.27	-0.16	0.36	-0.25	-0.27	0.50	
Stadsbus_Freq	0.44	0.45	0.48	0.14	0.40	0.32	0.06	0.10	0.14	0.17	0.30	0.18	0.32	0.06	0.73	0.31	0.67	1.00	0.11	0.57	0.42	0.02	0.30	0.98	0.02	0.57	0.41	0.30	0.91	0.41	0.12	0.17	0.41	0.58	-0.39	0.36	-0.37	0.36	0.24	0.21	0.27	-0.35	0.44	0.58	0.43	0.38	0.39	0.06	0.42	0.22	-0.12	0.51	0.40	-0.15	0.35	-0.14	-0.08	0.46	-0.16	-0.25	0.47	
Streekbus_Freq	0.58	0.50	0.50	0.42	0.29	0.37	0.28	0.08	0.35	0.37	0.40	0.16	0.16	0.15	0.24	0.34	0.77	0.11	1.00	-0.03	0.09	0.14	0.21	0.11	0.84	-0.04	0.09	0.83	0.08	0.22	0.22	0.34	0.24	0.28	-0.21	0.26	-0.26	0.23	0.20	0.17	0.26	-0.23	0.28	0.13	0.16	0.23	0.22	0.07	0.14	0.27	-0.25	-0.16	0.03	-0.21	-0.16	0.22						
Tram_Freq	0.37	0.42	0.31	0.11	0.13	0.10	0.06	-0.07	0.05	0.01	0.10	0.11	0.31	0.06	0.27	0.45	0.48	0.57	-0.03	1.00	0.24	0.07	0.15	0.57	-0.06	0.94	0.26	0.17	0.78	0.30	0.16	0.14	0.42	-0.42	-0.33	0.29	-0.33	0.25	0.16	0.26	-0.23	0.28	0.18	0.61	0.47	0.37	0.22	-0.03	0.50	0.37	-0.10	0.48	0.45	-0.11	0.47	-0.11	-0.05	0.50	-0.12	0.18	0.50	
Metro_Freq	0.33	0.26	0.46	0.18	0.30	0.16	0.10	-0.06	0.11	0.15	0.15	0.10	0.19	0.14	0.10	0.18	0.47	0.42	0.09	0.24	1.00	0.09	0.11	0.39	0.02	0.19	0.99	0.21	0.63	0.31	0.01	0.07	0.15	0.42	-0.32	0.20	-0.26	0.16	0.10	0.15	0.25	-0.30	0.51	0.34	0.29	0.27	0.24	0.06	0.26	0.28	0.06	0.23	0.22	-0.07	0.04	-0.06	0.04	0.23	-0.09	-0.04	0.24	
IC_NOL	0.24	0.23	0.22	0.10	-																																																									

APPENDIX 6: CORRELATION OF FINAL REGRESSION MODELS (MINUS OUTLIERS)

General Basic	General Basic													
	Daily_recent	Tot_Potentie	IC_Freq	Sprinter_Freq	BTM_NOL	CCI_2013_Actual								
	Daily_recent	1,00	0,81	0,35	0,39	0,57	0,60							
	Tot_Potentie	0,81	1,00	0,27	0,19	0,66	0,33							
	IC_Freq	0,35	0,27	1,00	-0,42	0,30	0,03							
	Sprinter_Freq	0,39	0,19	-0,42	1,00	0,01	0,50							
	BTM_NOL	0,57	0,66	0,30	0,01	1,00	0,06							
	CCI_2013_Actual	0,60	0,33	0,03	0,50	0,06	1,00							
General Extensive	Daily_recent	IC_Freq	Sprinter_Freq	CCI_2013_Actual	Streekbus_NOL	Stadsverv_NOL	Stadsbus_NOL	Parking_spaces	Bicycle_parking	Pot_Inw	Pot_MBO	Pot_HBO		
	Daily_recent	1,00	0,36	0,39	0,60	0,48	0,32	0,27	0,58	0,67	0,77	0,39	0,20	
	IC_Freq	0,36	1,00	-0,40	0,05	0,30	0,12	-0,05	0,32	0,29	0,26	0,18	-0,02	
	Sprinter_Freq	0,39	-0,40	1,00	0,51	-0,07	0,21	0,29	0,14	0,15	0,18	0,04	0,14	
	CCI_2013_Actual	0,60	0,05	0,51	1,00	-0,04	0,29	0,33	0,37	0,32	0,34	0,03	0,04	
	Streekbus_NOL	0,48	0,30	-0,07	-0,04	1,00	-0,07	-0,02	0,37	0,45	0,53	0,43	0,12	
	Stadsverv_NOL	0,32	0,12	0,21	0,29	-0,07	1,00	0,57	0,01	0,03	0,14	0,07	0,03	
	Stadsbus_NOL	0,27	-0,05	0,29	0,33	-0,02	0,57	1,00	0,11	0,23	0,30	0,06	0,02	
	Parking_spaces	0,58	0,32	0,14	0,37	0,37	0,01	0,11	1,00	0,47	0,46	0,24	0,03	
	Bicycle_parking	0,67	0,29	0,15	0,32	0,45	0,03	0,23	0,47	1,00	0,68	0,29	0,03	
	Pot_Inw	0,77	0,26	0,18	0,34	0,53	0,14	0,30	0,46	0,68	1,00	0,37	0,08	
	Pot_MBO	0,39	0,18	0,04	0,03	0,43	0,07	0,06	0,24	0,29	0,37	1,00	0,14	
	Pot_HBO	0,20	-0,02	0,14	0,04	0,12	0,03	0,02	0,03	0,03	0,08	0,14	1,00	
Regional Basic	Daily_recent	Tot_Potentie	Freq_Tot	Proximity										
	Daily_recent	1,00	0,85	0,43	-0,20									
	Tot_Potentie	0,85	1,00	0,32	-0,38									
	Freq_Tot	0,43	0,32	1,00	-0,23									
	Proximity	-0,20	-0,38	-0,23	1,00									
Regional Extensive	Daily_recent	Freq_Tot	Proximity	Pot_Inw	Pot_MBO	Pot_HBO								
	Daily_recent	1,00	0,43	-0,20	0,84	0,45	0,26							
	Freq_Tot	0,43	1,00	-0,23	0,25	0,18	0,19							
	Proximity	-0,20	-0,23	1,00	-0,32	-0,19	-0,18							
	Pot_Inw	0,84	0,25	-0,32	1,00	0,27	0,12							
	Pot_MBO	0,45	0,18	-0,19	0,27	1,00	0,04							
	Pot_HBO	0,26	0,19	-0,18	0,12	0,04	1,00							
Main Line Basic	Daily_recent	Tot_Potentie	BTM_NOL	CCI_2013_Actual	Freq_Tot									
	Daily_recent	1,00	0,78	0,58	0,47	0,62								
	Tot_Potentie	0,78	1,00	0,64	0,17	0,32								
	BTM_NOL	0,58	0,64	1,00	-0,03	0,17								
	CCI_2013_Actual	0,47	0,17	-0,03	1,00	0,56								
	Freq_Tot	0,62	0,32	0,17	0,56	1,00								
Main Line Extensive	Daily_recent	CCI_2013_Actual	Pot_Onderwijs	Pot_Joblnw	Streekbus_NOL	Stadsbus_NOL	Tram_NOL	Sprinter_Freq	IC_Freq					
	Daily_recent	1,00	0,49	0,39	0,73	0,51	0,22	0,31	0,39	0,29				
	CCI_2013_Actual	0,49	1,00	0,00	0,18	-0,12	0,28	0,27	0,62	-0,17				
	Pot_Onderwijs	0,39	0,00	1,00	0,36	0,27	0,07	0,08	0,12	0,07				
	Pot_Joblnw	0,73	0,18	0,36	1,00	0,56	0,27	0,15	0,14	0,21				
	Streekbus_NOL	0,51	-0,12	0,27	0,56	1,00	-0,05	-0,08	-0,13	0,34				
	Stadsbus_NOL	0,22	0,28	0,07	0,27	-0,05	1,00	0,57	0,31	-0,12				
	Tram_NOL	0,31	0,27	0,08	0,15	-0,08	0,57	1,00	0,22	0,07				
	Sprinter_Freq	0,39	0,62	0,12	0,14	-0,13	0,31	0,22	1,00	-0,48				
	IC_Freq	0,29	-0,17	0,07	0,21	0,34	-0,12	0,07	-0,48	1,00				

APPENDIX 7: OVERVIEW OF ALL STATIONS WITH ACTUAL AND ESTIMATED DEMAND.

Name	Actual	General Basic	General Extensive	Specific Basic	Specific Extensive	General Basic GWR	General extensive GWR	Specific Basic GWR
Aalten	1341	1339	1594	1641	#N/A	1320	1423	1609
Abcoude	1625	3396	3390	3304	787	3415	3322	3280
Akkrum	719	1035	1063	926	1060	1143	968	998
Almelo de Riet	1242	1833	1759	2075	584	2024	1848	2311
Almere Buiten	7900	5863	6662	6360	2824	6272	8777	6823
Almere Muziekwijk	7030	5366	5503	5683	5794	5723	6200	6050
Almere Oostvaarders	4285	4003	4128	4390	4184	4422	4428	4861
Almere Parkwijk	3907	4403	4688	4894	1201	4868	4965	5412
Alphen aan den Rijn	10130	8932	8379	8834	8860	9061	8439	8776
Amersfoort Schothorst	5642	6045	5637	5850	6190	6516	5939	6202
Amsterdam Muiderpoort	11147	11540	12257	12095	12100	11962	11972	12479
Anna Paulowna	2333	2014	2094	1531	2005	2183	2222	1607
Appingedam	1106	1374	1552	1100	1122	1415	1405	1151
Arkel	402	653	686	445	#N/A	531	490	433
Arnhemuiden	488	1163	960	586	865	1116	618	448
Arnhem Presikhaaf	3162	2528	2448	2743	3037	2457	2073	2887
Arnhem Velperpoort	3672	3708	3691	4355	4255	4177	3634	4886
Assen	9229	7561	7352	7947	8214	8019	7775	8699
Baarn	4658	3785	3997	3872	3879	3804	4513	3867
Baflo	666	151	101	529	1659	213	251	519
Barendrecht	4973	4939	3167	5244	2973	5023	3518	5259
Barneveld Centrum	3010	3491	3488	2772	606	3676	3233	2928
Barneveld Noord	1231	1820	1450	1138	#N/A	1899	1481	1288
Bedum	483	400	491	766	#N/A	405	544	797
Beek-Elstloo	2258	1766	1890	2062	509	1707	1849	2060
Beesd	183	555	408	538	2819	390	258	524
Beilen	2064	1451	1675	1653	1562	1538	2014	1755
Bergen op Zoom	7220	7927	8147	8051	8198	8225	8218	8050
Best	5322	4499	4674	4911	72	4651	4591	5078
Beverwijk	6237	5201	5181	5543	298	5207	5554	5569
Bilthoven	4380	4734	4551	4794	4684	4770	4951	4809
Blerick	1101	3431	3272	3159	3326	3840	3318	3482
Bloemendaal	1385	1898	1818	1773	1760	1681	1545	1543
Bodegraven	3005	3532	3780	2936	3479	3458	3278	2759
Borne	2348	2209	2604	2488	884	2412	2716	2753
Boskoop	1428	1545	1592	901	#N/A	1317	1272	894
Bovenkarspel Flora	867	1959	1658	1200	1702	2153	1572	1272
Bovenkarspel-Grootebroek	2399	2957	2892	2296	2663	3162	3180	2387
Boxmeer	4093	3761	3519	3081	#N/A	4002	4282	3233
Boxtel	6325	4135	4506	4443	4323	4244	6131	4544
Breda Prinsenbeek	1260	1819	1893	1886	1018	1590	1545	1672
Breukelen	5058	5160	5205	5375	5626	5561	5241	5730
Brummen	1075	987	1159	1119	1017	888	943	976
Buitenpost	1941	1916	1979	1342	95	2127	1906	1302
Bunde	954	617	691	802	653	519	687	741
Bunnik	2005	2846	2777	2858	58	2905	2616	2873
Bussum Zuid	3907	3144	2941	3137	3316	3182	3510	3138
Capelle Schollevaar	2242	3185	3254	3224	1032	3170	3612	3205
Castricum	7011	6095	5871	5130	1248	6604	6752	5455
Chevreumont	586	445	511	624	#N/A	335	375	612
Coevorden	1866	1958	2138	1965	4020	2182	2117	1909
Cuijk	3497	2829	2911	2451	#N/A	2963	2672	2595
Daarlerveen	116	95	49	393	#N/A	55	78	400
Dalen	202	-314	-305	247	#N/A	-438	-305	310
Dalfsen	1533	1559	1716	1292	#N/A	1750	1803	1261
De Vink	2783	3270	3667	3382	3528	3344	3487	3444
Deinum	137	-166	-295	295	#N/A	-205	-206	334
Delden	904	708	615	937	251	669	554	944
Delft Zuid	4668	3758	3689	3793	1003	3798	3607	3833
Delfzijl	1162	943	1108	700	#N/A	984	1001	745
Delfzijl West	442	420	328	754	#N/A	438	414	799
Den Dolder	1942	3071	2921	3050	3275	3104	2784	3044
Den Haag Mariahoeve	2877	3620	3183	3723	3308	3700	4184	3790
Den Helder	4180	5403	4889	5092	5562	5423	4864	5026
Den Helder Zuid	1918	1950	1683	1400	1767	2069	1979	1421
Deurne	4703	3352	3683	3112	3240	3663	3672	3328
Deventer Colmschate	1646	1777	2053	1874	793	1718	1869	1759
Didam	1899	2259	2168	2059	#N/A	2437	2129	2241
Diemen	3423	4167	3930	4118	1413	4173	3521	4074

Diemen Zuid	3304	3683	3578	3844	3054	3831	3420	3975
Dieren	3848	3407	3426	3239	3385	3669	3095	3425
Doetinchem	3968	4432	4379	3517	#N/A	4774	4501	3734
Doetinchem De Huet	1213	1829	1618	1600	400	2017	1775	1803
Dordrecht Stadspolders	709	2091	2301	1605	1594	2103	2177	1737
Dordrecht Zuid	1241	2076	2064	2252	1020	2079	2301	2194
Driebergen-Zeist	9267	7836	7904	7935	8410	8560	8797	8526
Driehuis	974	1862	1947	1792	565	1629	1611	1559
Dronrijp	155	-122	-214	342	#N/A	-166	-118	382
Duiven	3865	3261	3505	2779	#N/A	3442	3245	2967
Echt	2356	1463	1809	1700	2188	1386	1633	1675
Ede Centrum	1084	1801	1897	1538	3115	1653	2000	1530
Eijsden	213	633	789	764	236	532	642	737
Eindhoven Beukenlaan	1938	3237	2797	3389	2050	3299	2648	3461
Elst	3863	4505	4714	4354	749	5095	5348	4803
Emmen	2436	4167	3619	3597	#N/A	4245	2970	3658
Emmen Zuid	698	536	507	928	2913	536	477	965
Enkhuizen	2604	2689	2611	2015	2417	2894	2977	2109
Enschede De Eschmarke	81	158	157	281	#N/A	119	171	327
Enschede Drienerlo	2976	1700	1554	1897	1993	1873	1744	2093
Ermelo	2904	2405	2468	2460	2147	2261	2257	2277
Etten-Leur	3449	4142	4743	3725	1106	4070	3818	3530
Eygelshoven	312	285	147	296	#N/A	188	74	287
Franeke	918	828	1024	1309	#N/A	809	983	1350
Geerdijk	93	-15	-132	384	#N/A	-51	-71	395
Geldermalsen (NS+Arriva)	5856	4049	4358	4524	600	4437	4861	4788
Geldrop	1555	2258	2396	2267	1510	2077	1947	2092
Geleen Oost	600	807	811	954	686	706	819	898
Geleen-Lutterade	1504	832	698	895	706	701	640	791
Gilze-Rijen	2616	1899	2336	1882	1798	1672	2216	1671
Glanerbrug	308	442	681	610	#N/A	417	610	663
Goes	7660	7061	6300	7590	602	7408	6680	7536
Goor	1599	1420	1472	1525	804	1402	1239	1539
Gorinchem	4113	5109	4918	4050	-126	5352	4611	4168
Gouda Goverwelle	2835	4254	4325	4549	4573	4719	4505	4880
Gramsbergen	289	-177	-152	124	#N/A	-315	-202	188
Grijskerk	840	185	124	608	1598	181	199	630
Groningen Noord	1701	3603	3983	2891	#N/A	3880	3573	2780
Grou-Jimsum	923	852	828	599	2768	923	780	580
Haarlem Spaarnwoude	3086	3671	3303	3761	3978	3763	3413	3820
Harde 't	1236	902	859	944	1485	798	1096	772
Hardenberg	3175	3172	3434	3024	1274	3418	3254	3014
Harderwijk	5992	5167	5387	5702	5114	5203	5635	5793
Hardinxveld-Giessendam	660	1510	1714	1075	#N/A	1596	1654	1216
Haren	1132	1195	1247	1158	1867	1266	1531	1149
Harlingen	1840	1032	1222	1413	1774	1031	1102	1456
Harlingen Haven	341	366	430	669	#N/A	394	362	705
Heemskerk	2267	1939	2345	1928	1712	1690	2104	1691
Heemstede-Aerdenhout	6222	8410	7939	7570	8034	8878	8185	7767
Heerenveen	5782	5717	5901	6050	6395	6077	6885	6590
Heerhugowaard	7818	6508	7027	5920	6329	6714	7552	6028
Heeze	1634	1433	1796	1588	3100	1307	1588	1457
Heiloo	4614	4233	4556	3771	4218	4366	4400	3803
Heino	710	666	676	640	#N/A	589	656	646
Helmond	6847	6972	6956	7061	7172	7387	7347	7469
Helmond Brouwhuis	2057	1380	1379	1414	3244	1223	1182	1253
Helmond 't Hout	1247	1380	1327	1373	1299	1199	999	1168
Hemmen-Dodewaard	141	134	26	332	543	-34	-40	311
Hengelo Oost	2500	682	695	732	#N/A	651	671	781
Hertogenbosch 's Oost	1764	2325	1669	2391	1142	2150	1468	2219
Hillegom	2429	2746	2845	2763	2798	2570	2735	2577
Hilversum Noord	3795	2752	2437	2750	2884	2787	2243	2754
Hilversum Sportpark	7208	6202	6086	6459	739	6282	7371	6523
Hindeloopen	115	-455	-575	9	#N/A	-548	-513	68
Hoek van Holland Haven	1569	1293	1465	1462	1323	1270	1502	1421
Hoek van Holland Strand	494	-112	-167	-223	300	-493	790	-524
Hoensbroek	196	189	-47	264	36	68	-44	176
Hollandsche Rading	852	2280	1957	2050	2326	2050	1709	1787
Holten	1290	1237	1107	1443	1021	1192	1041	1366
Hoogeveen	4328	3999	3797	4439	4340	4126	5207	4661
Hoogezand-Sappemeer	1272	1716	1739	1780	369	1967	1876	1647
Hoogkarspel	2300	2442	2393	1714	2022	2638	2632	1785

Hoon Kersenboogerd	5388	4469	4290	3894	4353	4852	3940	4205
Horst-Sevenum	2635	1511	1470	956	1192	1533	1887	794
Houten	7478	5351	5201	5490	5233	5389	5261	5535
Houten Castellum	3499	4401	4569	4369	4525	4380	4983	4345
Houthem-St. Gerlach	341	284	221	398	#N/A	172	138	371
Hurdegaryp	1001	761	955	748	#N/A	802	850	772
IJlst	260	123	237	240	31	87	157	308
Kampen	4256	2975	3870	2807	#N/A	2961	4528	2858
Kapelle-Biezellinge	995	1501	1456	918	1171	1450	1121	767
Kerkrade Centrum	1115	774	940	894	#N/A	692	842	881
Kesteren	505	804	656	904	#N/A	681	469	884
Klimmen-Ransdaal	373	3	-80	170	943	-125	-56	154
Koog Bloemwijk	3016	3002	2993	2945	3145	2998	2783	2908
Koog-Zaandijk	3072	2612	2518	2619	1958	2651	2718	2637
Koudum-Molkwerum	160	-459	-557	58	#N/A	-564	-507	116
Krabbendijke	578	1626	1495	1228	1491	1695	1209	1144
Krommenie-Assendelft	5640	3570	3646	3700	3451	3593	3596	3716
Kropswolde	529	1192	1013	1105	522	1437	1284	970
Kruiningen-Yerseke	874	1242	1078	773	1127	1281	1105	678
Lage Zwaluwe	767	1471	1332	1677	1614	1530	1818	1683
Landgraaf	1228	1361	1258	1238	1362	1656	1526	1519
Leerdam	712	2615	2920	2343	180	2528	2485	2316
Leeuwarden Camminghaburen	835	1057	744	1175	#N/A	1026	730	1224
Leiden Lammenschans	3643	5179	5141	4962	2305	5445	4870	5055
Lichtenvoorde-Groenlo	878	681	652	857	#N/A	638	652	845
Lochem	1286	618	764	890	193	558	680	883
Loppersum	602	414	537	624	831	448	581	642
Lunteren	1044	1048	1334	710	942	879	1033	703
Maarn	1507	2705	2801	2758	2811	2758	2734	2752
Maarsse	4744	5946	5816	6197	6235	6109	6065	6341
Maassluis	2099	2269	1991	2489	1741	2324	1634	2492
Maassluis West	2510	2358	2020	2592	857	2403	1501	2582
Maastricht Randwyck	3672	2735	2385	3248	688	3184	2644	3880
Mantgum	495	540	529	682	1073	671	673	641
Marl�enberg	333	948	834	794	-37	1130	1065	767
Martenshoek	1013	1746	1732	1728	970	1991	1866	1596
Meersse	1223	1380	1262	1212	#N/A	1554	1468	1380
Meppel	5346	4883	4874	4924	5026	5260	6176	5396
Middelburg	4800	4846	5311	5062	1009	5095	5667	4994
Naarden-Bussum	9778	7390	7565	7835	7665	7737	7800	8164
Nieuw Amsterdam	784	295	329	633	#N/A	287	405	667
Nieuw Vennep	2556	3927	4357	4060	4181	4004	4834	4108
Nieuwerkerk a/d IJssel	3306	3335	3762	3385	676	3311	3641	3352
Nieuweschanse	588	4	-41	172	#N/A	45	109	197
Nijkerk	3650	2808	3384	3012	2991	2722	2926	2929
Nijmegen Dukenburg	2151	1745	1545	1748	786	1583	1232	1553
Nijmegen Heyendaal	3287	5387	4808	4666	390	5648	4634	4823
Nijmegen Lent	925	2775	2570	2433	2811	3083	2663	2573
Nijverdal	2939	2005	2326	2228	2624	1989	2028	2261
Nunspeet	2824	2113	2349	2340	2248	2055	2205	2274
Nuth	453	438	497	510	371	314	695	418
Obdam	1558	1216	1434	1406	1349	1170	1341	1366
Oisterwijk	2369	2299	2623	2450	151	2151	2435	2306
Oldenzaal	3275	3081	2669	3190	#N/A	3245	2484	3283
Olst	1293	1336	1400	834	1139	1299	1194	603
Ommen	1910	1402	1498	1310	824	1601	1589	1276
Oosterbeek	482	1037	1000	1097	1045	879	775	892
Opheusden	391	552	529	734	#N/A	400	338	714
Oss	8606	7679	7044	7631	1385	8021	8000	7964
Oss West	1806	1180	1363	1162	1100	987	1193	928
Oudenbosch	1257	1989	1978	2217	1811	1839	1541	2056
Overveen	3092	1768	1710	1648	1863	1557	1371	1428
Purmerend	2992	3694	3710	3759	3818	3481	3438	3520
Purmerend Overwhere	2312	4003	4231	4159	4195	3823	3898	3961
Putten	1914	1456	1721	1461	241	1288	1707	1256
Raalte	2059	2158	2394	2227	#N/A	2140	2391	2257
Ravenstein	1364	892	974	935	858	726	963	723
Reuver	1519	995	1266	1194	#N/A	886	1034	1162
Rheden	907	1131	1330	1220	617	1000	1007	1042
Rhenen	1401	1739	2039	1963	1929	1671	1780	1888
Rijssen	2428	2769	3238	3027	2844	2775	3075	3029
Rijswijk	7141	8162	6672	8530	7174	8184	6413	8469
Rilland-Bath	437	1111	836	537	888	1108	601	420
























Roodeschool	247	-198	-290	531	#N/A	-175	-127	545
Rosmalen	2324	2837	2964	2867	2608	2644	2468	2689
Rotterdam Lombardijen	6272	6025	6319	6726	6002	6346	5366	6905
Rotterdam Noord	2302	2783	3778	2974	2853	2876	2995	3031
Rotterdam Zuid	2799	3379	3008	3493	669	3445	2824	3530
Ruurlo	951	539	706	748	#N/A	482	589	716
Santpoort Noord	864	1431	1518	1337	2762	1233	1272	1134
Santpoort Zuid	866	1363	1375	1262	1332	1171	1155	1065
Sappemeer Oost	551	1163	1012	1217	575	1417	1290	1081
Sauwerd	372	971	894	952	798	1219	1175	801
Schagen	5921	4281	4270	3833	4008	4505	4073	3956
Scheemda	694	436	546	487	117	501	649	495
Schiedam Nieuwland	4835	3303	4648	3669	4634	3602	3814	3851
Schin op Geul	371	-6	-126	150	1516	-136	-110	134
Schinnen	346	216	169	316	133	96	188	224
Sliedrecht	866	1542	1413	797	#N/A	1659	1650	939
Sneek	2901	4196	4270	3178	#N/A	4488	3847	3175
Sneek Noord	959	773	737	1140	#N/A	852	833	1116
Soest	231	955	873	891	556	749	603	661
Soest Zuid	2060	1927	2278	2052	2135	1817	2002	1941
Soestdijk	938	1457	1348	1485	1387	1275	1026	1293
Spaubeek	397	172	23	238	38	42	-15	134
Stavoren	331	-480	-569	39	207	-587	-505	95
Stedum	287	-126	-237	425	#N/A	-114	-91	443
Steenwijk	3021	3259	3399	3080	3274	3454	4043	3244
Susteren	928	586	712	678	502	461	592	572
Swalmen	453	615	700	494	#N/A	508	554	486
Tegelen	745	1209	1319	1288	#N/A	1116	1186	1274
Terborg	711	863	956	1020	#N/A	795	816	988
Tiel	4128	5257	4893	5779	2670	5447	5665	5997
Tilburg Reeshof	2563	3141	3064	3226	2744	2944	2720	3040
Tilburg Universiteit	7348	5348	5428	5576	6238	5364	5911	5606
Utgeest	5336	4860	5392	5268	5252	5363	5582	5843
Uithuizen	891	383	502	902	#N/A	414	518	917
Uithuizerveeden	420	-86	-113	634	40	-65	14	648
Usquert	226	-127	-220	559	1580	-108	-87	573
Utrecht Lunetten	3458	2573	2481	2672	1371	2660	2787	2726
Utrecht Overvecht	6827	6460	6838	6969	870	7040	7460	7480
Utrecht Terwijde	2626	2232	2264	2081	2343	2042	1923	1904
Valkenburg	1691	1666	1816	1329	3121	1866	2202	1494
Varseveld	533	560	577	871	#N/A	493	506	837
Veenendaal Centrum	2438	5278	#N/A	#N/A	5404	5523	4881	6099
Veenendaal West	1549	3619	3392	3889	3584	3756	3269	4009
Veenendaal-de Klomp	3745	3140	2969	2764	893	3178	3637	2693
Veenwouden	982	401	493	521	#N/A	409	483	555
Velp	1625	1648	1620	1634	3878	1500	1549	1423
Venray	3295	2836	3128	2331	#N/A	3033	3132	2489
Vierlingsbeek	596	1025	1037	725	#N/A	1135	1083	886
Vlaardingen Centrum	2790	3720	3525	4120	3614	4165	3471	4413
Vlaardingen Oost	2817	4979	4190	5278	1092	5154	4685	5354
Vlaardingen West	1951	2514	1914	2653	2044	2533	2279	2638
Vleuten	3334	2812	3039	2724	2921	2637	2645	2566
Vlissingen	2999	2270	1826	1790	2140	2212	2970	1622
Vlissingen Souburg	1007	1856	1938	1457	1718	1876	1486	1340
Voerendaal	346	129	67	201	#N/A	17	56	184
Voorburg	2050	4195	1701	4573	1555	4289	2210	4598
Voorhout	3452	2778	3092	2726	2747	2544	3039	2501
Voorschoten	2917	3292	3541	3356	3544	3365	3343	3423
Vorden	1051	487	556	688	930	412	504	672
Vriezenveen	311	465	590	839	1097	429	544	853
Vroomshoop	341	421	494	838	-135	391	574	856
Vught	2010	2244	2346	2251	218	2044	2089	2051
Waddinxveen	1575	2013	2139	1313	275	1790	1692	1301
Waddinxveen Noord	1238	1657	1795	852	3250	1475	1499	845
Warffum	695	240	89	815	1040	306	242	793
Weert	7390	6927	7501	6560	-221	7868	9063	7356
Weesp	9440	7072	7007	7449	384	7703	7723	8041
Wehl	567	1452	1543	#N/A	577	1608	1602	1252
Wezep	1067	1172	1380	1266	1175	1091	1350	1120
Wierden	1837	1620	1849	1818	601	1798	2197	2022
Wijchen	4214	3388	3874	3628	687	3314	3335	3619
Wijhe	1125	1273	1201	769	1035	1242	1121	537
Winschoten	2447	2691	2505	2473	#N/A	2782	2273	2491
Winsum	2382	1027	1184	1088	428	1101	1126	1081
Winterswijk	1935	2106	2024	2048	#N/A	2353	2037	2276


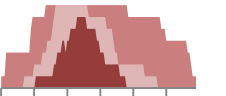
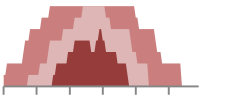
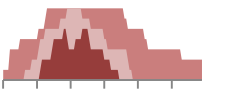

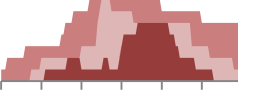
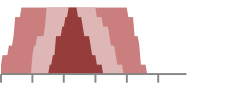
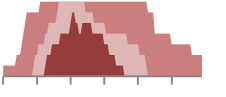
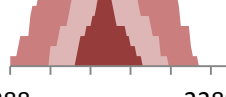
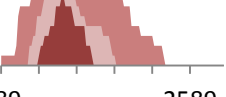
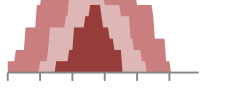
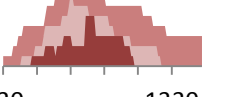

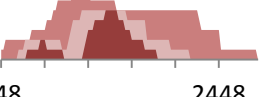

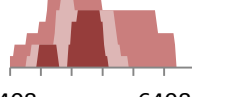
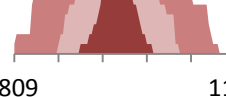
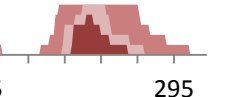
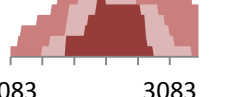
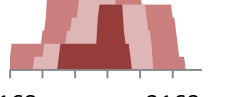
Winterswijk West	372	540	173	794	#N/A	502	170	813
Woerden	11648	8710	#N/A	8576	1984	9209	9096	8918
Wolfheze	542	887	743	944	1762	725	560	730
Wolvega	1521	2245	2503	2422	1019	2369	2400	2580
Workum	476	-142	-107	224	#N/A	-221	-133	288
Wormerveer	4091	3859	3751	4005	4097	3881	4148	4015
Zaandam Kogerveld	1713	2178	2169	2215	2279	2093	2019	2098
Zaltbommel	3417	2760	2559	3012	4007	2681	3594	2944
Zetten-Andelst	732	923	904	837	#N/A	794	690	819
Zevenaar	4652	2679	3203	2005	#N/A	2873	2942	2198
Zevenbergen	1263	1525	1595	1693	1405	1361	1552	1523
Zoetermeer	5947	7686	7144	8585	822	7979	7933	8714
Zoetermeer Oost	3196	2445	1870	2616	2176	2533	2189	2672
Zuidbroek	809	1230	1326	1082	1079	1491	1520	942
Zuidhorn	2595	1328	1642	1265	#N/A	1447	1640	1221
Zwaagwesteinde	614	332	468	641	#N/A	320	452	676
Zwijndrecht	5264	5961	6168	6272	6053	6000	6484	6251

APPENDIX 8: OVERVIEW OF ALL VALIDATION STATIONS AND THEIR ESTIMATED RIDERSHIP FOR ALL MODELS.

Name	Actual	Actual (year after)	General Basic	General Extensive	Regional Basic	Regional Extensive	Main_Line_basic	Main_Line_Extensive	General basic (GWR)	general extensive (GWR)	Regional basic (GWR)	regional extensive (GWR)
Sliedrecht Baanhoek	553		1640	1069	1564	1098	2701	1720	1495	925	1798	2685
Groningen Europapark	989		2549	2127	1535	1305	2908	2722	3078	2639	1462	3848
Hardinxveld Blauwe Zoom	246		601	456	531	349	631	392	396	222	542	402
Halfweg	1478	1487	3717	3561	1257	1289	3749	4059	3826	3752	1408	3804
Amsterdam Holendrecht	3176	3024	6326	7536	3281	2508	6724	5125	6549	3566	3483	6890
Gaanderen	339		548	507	583	540	609	413	456	412	622	452
Almere Poort	2256	2256	4466	3738	1773	1516	4792	4800	4891	4155	2128	5218
Apeldoorn De Maten	619		989	868	967	992	1032	793	885	756	1014	863
Den Haag Ypenburg	1801	908	2887	3929	1496	1843	3044	3105	2918	3333	1711	3049
Purmerend Weidevenne	1646	1644	2436	2542	1512	2100	2381	2386	2230	2386	1597	2143
Mook Molenhoek	1224		1708	1814	1262	1502	1898	1716	1785	1888	1481	1940
BovenHardinxveld	343		506	505	357	342	595	463	336	303	370	386
Apeldoorn Osseveld	1040	640	1257	1546	1141	1302	1284	1131	1149	1463	1191	1106
Arnhem Zuid	2790		3918	3707	1789	2266	3598	4070	4244	3971	2109	3781
Helmond Brandevoort	1021	744	1493	1075	1030	675	1482	1262	1308	888	1058	1275
Sassenheim	3000	3000	3858	4512	1851	1816	3966	3997	3905	4865	2013	3993
Voorst-Empe	342		386	329	383	224	495	367	285	204	368	337
Amersfoort Vathorst	2559	1132	3121	2996	1840	1903	3282	3118	3191	3107	2090	3334
Nijmegen Goffert	1000		1312	910	917	528	1313	1170	1145	725	952	1100
Veendam	2000		1818	1847	1856	2279	2231	1894	1857	1913	1898	2315
Twello	1554	1224	1483	1690	1208	1635	1689	1539	1420	1660	1222	1605
Eygelshoven Markt	285		242	232	398	343	325	93	133	111	430	252
Kampen Zuid	1141	1141	1135	1229	718	755	1274	1221	1094	1210	739	1194
Heerlen de Kessel	371		319	297	366	347	432	242	212	224	398	363
Klarenbeek	283		305	284	247	56	367	247	188	146	244	180
Barneveld Zuid	900		945	687	833	535	954	688	762	493	834	735
Tiel Passewaaij	1269	952	1204	1187	601	800	1222	1166	1021	991	635	998
Westervoort	2250		1823	1667	1229	1216	2007	1873	1909	1736	1474	2057
Hoevelaken	1500		1406	1158	447	235	1519	1454	1368	1088	596	1459
Heerlen Woonboulevard	85		9	-88	42	-252	76	-116	-118	-216	78	-29
Dronten	3142	3142	2032	2092	1609	2059	2301	2090	2023	2075	1638	2298
Maarheeze	1258	1176	695	814	513	435	829	728	568	717	495	705
Utrecht Leidsche Rijn	4700		2603	2332	465	294	2656	2812	2718	2425	768	2736
Hengelo Gezondheidspark	1450		541	318	558	222	603	381	497	258	646	468

APPENDIX 9: SELECTION OF PROPOSED STATIONS WITH RIDERSHIP ESTIMATION AND ERROR MARGINS

Name	Estimation	Error margin (absolute)	Error Margin (relative)	Name	Estimation	Error margin (absolute)	Error Margin (relative)
's-Hertogenbosch Maaspoort	Rel: 1483			Maartens dijk	Rel: 2598		
	Abs: 1471				Abs: 2535		
's-Hertogenbosch Avenue	1468			Nijkerk Corlaer	1287		
	1455				1314		
Apeldoorn West	2118			Oss Oost	869		
	2090				903		
Arnhems Buiten	909			Ressen	943		
	884				849		
Baexem	825			Schiedam Kethel	3561		
	776				3739		
Belfeld	431			Sneek-Harinxmal and	N.A.		N.A.
	411				-88		

Berkel Enschoot	1807			Stadskanaal (centrum)	1574		
	1844	594 2594	837 2837		1613	443 2443	684 2684
Breda Oost	2883			Staphorst	827		
	2345	655 2655	873 2873		835	-75 1925	407 1407
Deventer Platvoet	1239			Stroe	845		
	1228	288 2288	589 2589		829	-251 1749	330 1330
Deventer Zuid	1458			Utrecht Lage Weide	3958		
	1446	-54 1946	448 2448		3885	1760 4260	1408 6408
Duurkenakker	155			Utrecht Majella	2408		
	61	-809 1191	-5 295		2223	1083 3083	1168 3168

LIST OF FIGURES & TABLES

LIST OF TABLES

Table 1: Overview of all estimated regression models.....	5
Table 2: Comparison between predicted and actual ridership demand.....	11
Table 3: Overview of all variables linked to ridership generation	20
Table 4: Overview of all data sources	41
Table 5: Station to be used in the validation phase.....	41
Table 6: Top 5 of best and worst scores for the CCI and SCI indices	46
Table 7: Overview of various intercity and sprinter stations and their corresponding SCI and CCI index scores.	47
Table 8: Basic statistics of the access distance to the station per station type and mode.....	50
Table 9: Basic statistics of the egress distance from the station per station type	51
Table 10 & 12: decay function per station type (left) and various tested function types (right).	53
Table 11: Number of observations per rank	57
Table 12: Overview of MNL station choice model results	59
Table 13: Overview of the model parameters of Station Choice model 2.	61
Table 14: Overview of the model parameters of station choice model 3.	62
Table 15: Correlation of independent variables with the dependent variable (Daily_2013).....	68
Table 16: Overview of all regression model results.....	75
Table 17: Stations with at least 3 times the standard error per model.	77
Table 18: Overview of all regression models after removal of outliers.....	78
Table 19: Model fit of the geo-weighted and the regular regression models	79
Table 20: Overview of model estimates	83
Table 21: Change of demand for stations in the vicinity of new stations opened between 2005 and 2013.....	84
Table 22: The demand change at existing stations near new stations as a result of changes in the CCI indicator values.....	85
Table 23: Total demand change modelled (potential and CCI combined)	86
Table 24: List of proposed stations with an estimation based on absolute and relative error margins.	90
Table 25: Overview of stations affected by the opening of a new station. This includes the decrease in demand by abstraction and as an effect of a reduced rail accessibility.	91

LIST OF FIGURES

Figure 1: Distance decay functions per station type on the access side of the trip	4
Figure 2: demand abstractio of Leeuwarden as a result of the opening of Leeuwarden-Werpsterhoek	5
Figure 3: The balance of a new station	21
Figure 4: The effects that can be expected when adding a new station	33
Figure 5: Conceptual model for ridership estimation.....	34
Figure 6: Overview of the most important direct connections of the two sprinter stations in Apeldoorn (right).	43
Figure 7: Trip distribution in actual journey time.....	45
Figure 8: CC index plotted against the SC index	46
Figure 9: Graphs of the distance decay function per station type	55
Figure 10: Distance decay functions for sprinter, intercity stations & for egress.	56
Figure 11: Share of pedestrians, cyclists, and public transport users plotted against the distance to the station. Share of car users and other modes excluded. Source: Stedenbaan survey.	58
Figure 12: Change in demand after the opening of Leeuwarden Werpsterhoek as modelled with (from left to right) choice model 1, model 2 and, model 3.	64
Figure 13: Actual versus Potential ridership demand measured in daily boarding per station	66
Figure 16: Basic (left) and extensive (right) Main Line model scatterplots (estimated vs actual ridership).	76
Figure 14: Basic (left) and extensive (right) general model scatterplots (estimated vs actual ridership)	76
Figure 15: Basic (left) and extensive (right) regional Models scatterplots (estimated vs actual ridership)	76
Figure 17: Absolute error plotted against the percentage of cases that falls within the error margin ...	87
Figure 18: Relative estimation error plotted against the percentage of cases that falls within the error margin.....	88
Figure 19: Relative Error margins for Station of Meppel with a 25, 50 and 75 percent certainty.....	89
Figure 20: Absolute Error margins for Station of Meppel with a 25, 50 and 75 percent certainty.....	89
Figure 21: Total balance for the stations of Gorinchem Noord (left) and Oss Oost (right).....	92