Determinants of the demand for bus transport

A model to determine the number of bus boardings in a neighbourhood

Erik Klok

regio twente

UNIVERSITY OF TWENTE.

Determinants of the demand for bus transport

A model to determine the number of bus boardings in a neighbourhood

Enschede, August 2010

Master Thesis of:
E.J. Klok
Centre for Transport Studies
Civil Engineering & Management
University of Twente

Supervisors:

Dr. Ing. K.T. Geurs
Centre for Transport Studies
University of Twente

Dr. T. Thomas
Centre for Transport Studies
University of Twente

Drs. M.L. Berloth
Department Mobility
Regio Twente

UNIVERSITY OF TWENTE.

regio twente

Index

i Summary iii
ii Samenvatting vi
iii Preface xi
iv Dankwoord xi
1 Introduction 1
2 Problem description 3
2.1 Research background 3
2.2 Research objective 4
2.3 Research questions 4
3 Literature study 5
3.1 Possible determinants 5
3.2 Travel demand modelling 12
4 Research methodology 17
4.1 Modelling dependence 17
4.2 Spatial resolution 18
4.3 Project scope 19
4.4 Model 21
4.5 Time horizon 22
5 Research Area 23
5.1 Location in Twente 23
5.2 Enschede 23
5.3 Haaksbergen 24
5.4 Losser 25
6 Zoning 26
6.1 Subdivision 26
6.2 Neighbourhoods serviced by multiple lines 27
7 Data 29
7.1 Overview 29
7.2 Sources 29
7.3 Boardings per Area/Stop (per inhabitant) 34
7.4 Socio economic 37
7.5 Built environment 39
7.6 Transportation network 42
7.7 Conclusion 44
8 Correlation 45

8.1 Dependent variable 45
8.2 Independent variables 48
8.3 Conclusion 54
9 Regression 56
9.1 Design 56
9.2 Initial regression 56
9.3 Next steps 57
9.4 Best model(s) 60
9.5 Including the university and the harbour area 64
9.6 Conclusion 68
10 Validation 69
10.1 Model outcomes 69
10.2 Validation with 2004 data 73
10.3 Conclusion 79
11 Conclusion 80
11.1 Answering research questions 80
11.2 Definitive model 81
11.3 Comparing to literature 82
11.4 Recommendations & Further research 84
Bibliography 86
Appendix A: Variables and their interdependencies 88
Appendix B: Subdivision into zones 91
Appendix C: Relative accessibility 93
Appendix D: Distance to stop 99
Appendix E: Correlation in urban and rural areas 102
Appendix F: Correlation between independent variables in Neighbourhoods 103
Appendix G: Correlation between variables in zones (research objects) 104
Appendix H: Regression 105

i Summary

i.1 Research objective

Regio Twente, the bus concession granting authority in 14 eastern municipalities in the Netherlands, is interested to know what the most important determinants are of the demand for bus transport. The objective of this study therefore is to develop a model that describes the relations between the demand for bus transport and its most important determinants.

i.2 Research design

The first part of the report described the choices made on the research design. The choices considered the dependent variable, the research area, the variables to study and the research methodology.

i.2.1 Dependent variable definition

The dependent variable of the research is the demand for bus transport. The measure for the demand of bus transport is the number of boardings per inhabitant on an average workday in 2008. This measure is chosen because it is the only measure with a reliable source (the NVS-counts) that covers the whole research area. Only the number of passengers at certain stops is counted. Therefore the net number of passengers is assumed to be the number of boardings in a certain zone.

i.2.2 Research area and subdivision into research objects

To reduce the effort that has to be done on obtaining the data, the research area is limited to three municipalities in Twente that all have different characteristics: Enschede (urban), Haaksbergen (rural, many lines) and Losser (rural, few lines).
The research area is subdivided into zones (research objects) that are formed by groups of neighbourhoods. This scale level is used because the boarding counts can only be attributed to groups of neighbourhoods rather than to neighbourhoods alone. In the end the research area is subdivided into 23 zones, of which some overlap each other because some neighbourhoods are serviced by more than one line with different destinations/routes.

i.2.3 Variables included

The variables included in the research, their subdivision into socio economic, built environment and network variables and their mutual relations are presented in figure i.1.

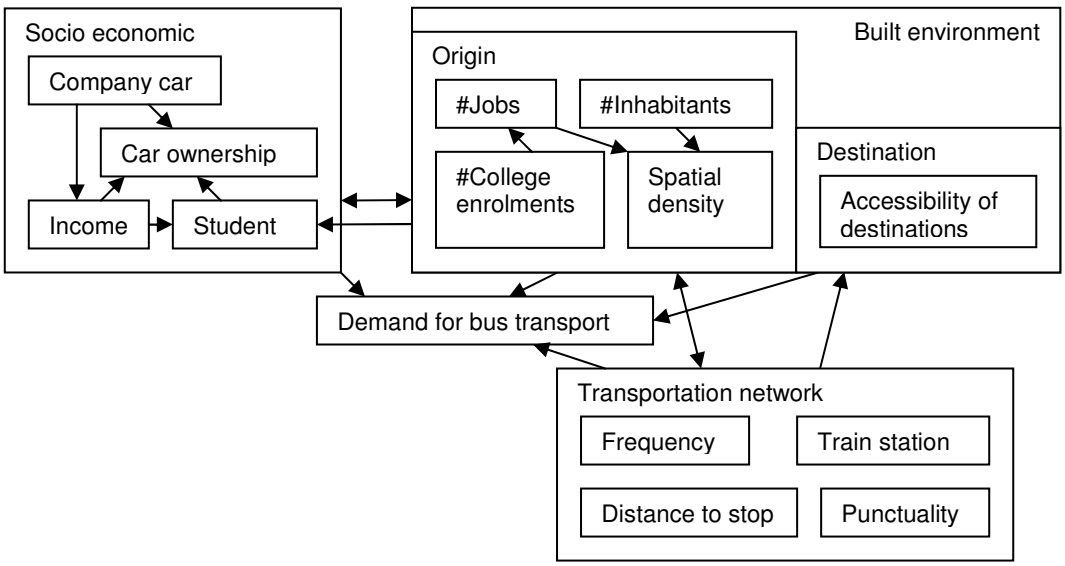


Figure i.1 Variables included in the research and their mutual relations

i.2.4 Research methodology

The research itself contains three analytical steps: The correlation, regression and validation. In the first step the correlation between the different (dependent and independent) variables are computed to show whether or not the variables should be included in a model of the demand for bus transport. In the regression different linear models are estimated that all are possible models and in the validation these models are compared and the best model is chosen.

i.3 Analysis

i.3.1 Correlation

In the correlation analysis first the correlation of the independent variables with the dependent variable is studied. Variables with low correlation will not improve the model when included. Also the mutual correlation coefficients of the different independent variables are studied to check which variables can and cannot be used in a regression model together.

i.3.2 Regression

During the regression a total of 26 models were estimated.
For the first 18 models three zones are not included (the city centre of Enschede, the university and the harbour of Enschede) because extreme values of the number of boardings per inhabitant or the number of jobs per inhabitant were present in these zones. From these models two models are indicated that are possibly best.
For the other eight models the university and/or harbour zone are included. In two cases the number of boardings in the university zone is adjusted to correct for the number of students not living on the university campus that board the busses. From these models only one model, with both unadjusted zones included, is analysed further.

i.3.3 Validation

The three potential models resulting from the regression phase are then validated. This validation took place with both data of the study year (2008) and data of a control year (2004).
From the 2008 data is can be concluded that the number of boardings is slightly underestimated for large numbers of boardings and overestimated for low numbers of boardings. No other variables however showed a relation with the residuals of the models which indicates that including another variable is does not improve the models.
The under- and overestimation of the model with the 2004 data is larger than for the 2008 dataset. Between 2004 and 2008 a major frequency increase was present, but the numbers of boardings did not change considerably. This probably is the result of the fact that the counts are only made during a limited period every year, which makes weather influences possible. Also the introduction of competitive lines may have led to lower numbers of boardings on the main line. It is also possible that a certain ceiling for the number of boardings per inhabitant is present. Increasing the frequency when this ceiling is reached then does not lead to extra travellers. It also seems that the number of jobs per inhabitant is less important in 2004 than in 2008.

i.4 The model

The model parameters of the best model according to the validation are presented in table i.1.

Model (R^2 = 0.804)	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta	B	Std. Error
Constant	-0.117	0.036		-3.230	0.006
% of people aged 15-24	0.00161	0.00133	0.183	1.212	0.246
Address density	-2.46E-05	6.70E-06	-0.593	-3.669	0.003
Frequency	0.00099	0.00013	0.852	7.869	0.000
Jobs per inhabitant	0.0229	0.0134	0.188	1.716	0.108
Punctuality	0.00185	0.00054	0.411	3.441	0.004

Table i.1 Model parameters

In the table the parameter values (column 2) and their standard errors (column 3) are given. Also the standardized coefficients (column 4), the t value (column 5) and the corresponding significance level (column 6) are included.
The standardised coefficients of the variables give an indication of the relative importance of the variables. When a value of Beta is twice as high as another Beta the variable is twice as important. In table i.1 it can be seen that the frequency is the most important determinant of the demand for bus transport. The importance of the address density is about 70% of the importance of the frequency and the other variables are 20% (students and jobs) and 50% (punctuality) as important as the frequency.

i.5 Recommendations

Some questions regarding the demand for bus transport can be answered with the outcomes of this study. There however also are questions that remain with possibilities for future research, like:

- Using (reliable) PT-chipcard boarding numbers.
- Using multiple years to estimate a model.
- Develop a better measure for the relative attractiveness of bus transport.
- Study how major attraction points (like the city centre of Enschede) can be included in the model.
- Including other variables, like; land use, trip purpose, personal preferences and household composition.

ii Samenvatting

In deze uitgebreide Nederlandse samenvatting komen de verschillende delen van het rapport en het bijbehorende onderzoek aan de orde en worden de uitkomsten van het onderzoek besproken.

ii.1 Aanleiding

Regio Twente, het samenwerkingsverband van 14 Twentse gemeenten, is de concessieverlenende autoriteit voor busvervoer in het oosten van Overijssel. In het kader van de nieuwe aanbesteding voor de concessieperiode vanaf eind 2013 wil zij graag weten waarvan de vraag naar busvervoer op wijk- of halteniveau afhangt. Daarom is het doel van dit afstudeeronderzoek een model te ontwikkelen dat beschrijft op welke manier verschillende netwerk-, ruimtelijke en persoonlijke karakteristieken de vraag naar busvervoer beïnvloeden.

ii.2 Onderzoeksopzet

De verschillende belangrijke keuzes en aannames aangaande het onderzoek worden hier besproken. Het gaat hierbij om keuzes met betrekking tot de afhankelijke variabele, het onderzoeksgebied, de te onderzoeken variabelen en de onderzoeksmethodologie.

ii.2.1 Definitie van de afhankelijke variabele

De vraag naar busvervoer is natuurlijk een nogal vaag begrip. In het onderzoek wordt daarom gebruik gemaakt van een maat voor de vraag naar busvervoer: het aantal instappers per inwoner op een gemiddelde werkdag in 2008. Het aantal instappers wordt hierbij bepaald door het netto aantal reizigers in de bus in een bepaalde zone. Het gaat hierbij om netto reizigers, aangezien de tellingen die de bron zijn van de gegevens (de NVS-tellingen) alleen het aantal reizigers op een bepaalde halte weergeeft. Wanneer er op een gemiddelde werkdag 600 mensen in de bus zitten voordat de zone wordt aangedaan en 1000 nadat de zone is aangedaan dan is het netto aantal reizigers voor de zone dus 400. Er wordt vervolgens aangenomen dat het aantal instappers in deze zone gelijk is aan het netto aantal reizigers. In werkelijkheid kunnen er ook mensen uitstappen in de zone, waardoor het netto aantal reizigers dus lager is dan het aantal instappers. Het aantal instappers kan dus alleen maar hoger zijn dan aangenomen. Aangezien het aantal instappers per inwoner wordt bekeken is het ook van belang te weten waar het aantal inwoners vandaan komt. Voor het onderzoek is gebruik gemaakt van de cijfers in het Regionale Verkeersmodel Twente (RVM Twente). Er kan echter ook gebruik worden gemaakt van de cijfers van het Centaal Bureau voor de Statistiek (CBS).

ii.2.2 Onderzoeksgebied en onderverdeling naar onderzoeksobjecten

Omdat het onmogelijk is om het hele concessiegebied te beschouwen is ervoor gekozen om het onderzoek te beperken tot drie gemeenten: Enschede, Haaksbergen en Losser. Op deze manier wordt er zowel een stedelijke gemeente (Enschede), een landelijke gemeente met veel buslijnen (Haaksbergen) als een landelijke gemeente met weinig buslijnen (Losser) meegenomen.
Deze gemeenten zijn vervolgens opgedeeld in 23 onderzoeksobjecten. Deze onderzoeksobjecten zijn zones (waarbij één zone een combinatie van verschillende CBS-buurten is) die allen worden bediend door een (aantal) lijn(en). Met behulp van NVS-tellingen is vervolgens het aantal instappers in de verschillende zones bepaald.

ii.2.3 Te onderzoeken variabelen

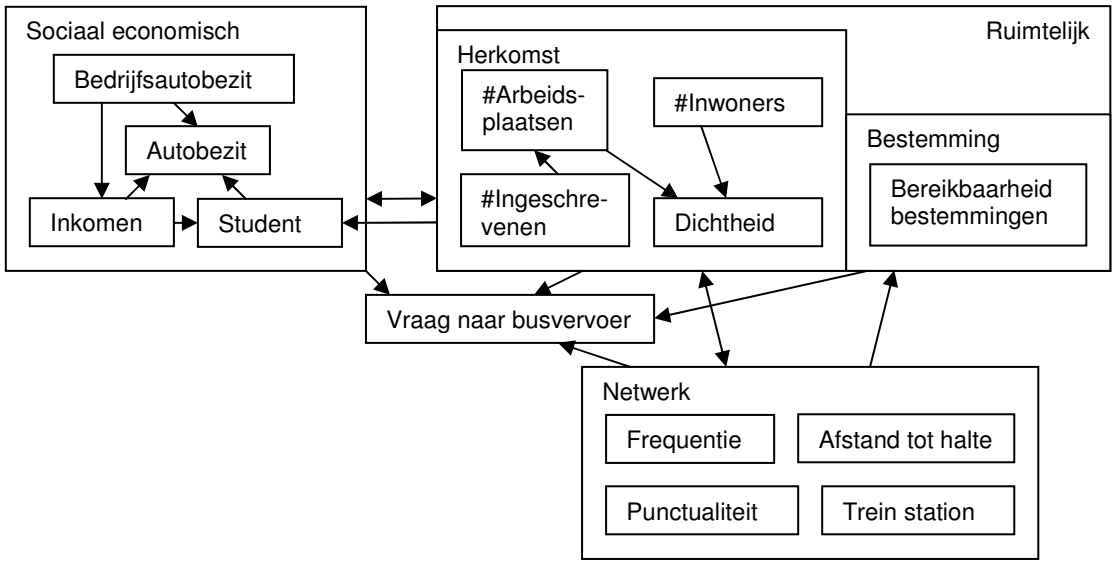
Na een uitgebreid literatuuronderzoek is ervoor gekozen een aantal verklarende variabelen (determinanten) voor de vraag naar busvervoer te onderzoeken die veel werden genoemd of waarvan een sterk vermoeden bestond dat het belangrijke determinanten zijn. Deze variabelen zijn[i]:

- Sociaal economische variabelen
    - Autobezit
    - Bedrijfsautobezit
    - Inkomen

[i] Tussen haakjes staan de maat voor de variabele als die anders is dan de naam van de variabele doet vermoeden.

    - Studenten (het percentage studenten en het percentage inwoners van 15 tot en met 24 jaar)
- Ruimtelijke variabelen
    - Inwoners
    - Arbeidsplaatsen
    - Bereikbaarheid van bestemmingen (relatieve arbeidsplaatsbereikbaarheid per auto of fiets ten opzichte van het openbaar vervoer, reistijd naar het centrum van Enschede per bus en de relatieve reistijd per auto en fiets naar het centrum van Enschede in vergelijking met de reistijd per bus)
    - Dichtheid (inwoner-, adressen-, en banendichtheid)
    - Het aantal ingeschreven studenten (WO, HBO, MBO)
- Netwerkvariabelen
    - Frequentie
    - Punctualiteit
    - Afstand tot een bushalte
    - De aanwezigheid van een (intercity) treinstation

De onderlinge relaties tussen deze variabelen zijn weergegeven in figuur ii.1.



Figuur ii.1 Relaties tussen de onderzochte variabelen

ii.2.4 Onderzoeksmethodologie

Het onderzoek zelf bestaat uit drie analysestappen. Als eerste is er een correlatieanalyse uitgevoerd, waarin er is gekeken welke variabelen wel en welke niet tegelijkertijd in een model kunnen worden gebruikt. Vervolgens is met behulp van lineaire regressie een aantal modellen geschat met wisselende samenstellingen van variabelen. In de validatie is tot slot bepaald welk model het beste gebruikt kan worden om de vraag naar busvervoer te schatten.

ii.3 Analyse

De analyse is het eigenlijke onderzoek. Dit is onderverdeeld in drie delen, de correlatieanalyse, regressie en validatie.

ii.3.1 Correlatieanalyse

Een correlatieanalyse is een manier om in kaart te krijgen welke variabelen wel en welke niet aan elkaar gerelateerd zijn. Wanneer er geen relatie bestaat tussen de afhankelijke variabele (hier: het aantal instappers per inwoner) en een andere variabele, dan heeft het geen zin om de variabele in een model voor de afhankelijke variabele te stoppen. Wanneer er een sterke correlatie bestaat tussen de verschillende potentiële onafhankelijke variabelen dan heeft het geen zin om beide variabelen in hetzelfde regressiemodel mee te nemen.

ii.3.2 Regressie

Met behulp van lineaire regressie zijn in totaal 26 modellen geschat.
Allereerst zijn er 18 modellen geschat waarbij drie zones niet meegenomen zijn. Dit zijn het centrum van Enschede, de universiteit en de haven van Enschede. In het centrum is het aantal instappers buitengewoon hoog doordat er veel overstappers van de trein komen. Op de universiteit stappen veel studenten in die daar niet wonen en in de haven zijn door het lage aantal inwoners extreem hoge aantallen arbeidsplaatsen per inwoner en een extreem hoog autobezit te vinden. Van deze modellen blijven uiteindelijk twee mogelijke modellen over die voldoen aan de criteria met betrekking tot de significantie van de parameters, de tekens voor de parameters en de achtergrond van de variabelen. Van elk van deze twee modellen worden vervolgens nog drie modellen geschat die zijn gebaseerd op dezelfde zones met daarbij de zone van de universiteit, de haven of beide. Aangezien het aantal instappers per inwoner in de universiteitszone extreem hoog is zijn er ook nog 2 modellen geschat met een aangepast aantal instappers in deze zone, dit leidt echter niet tot het gewenste significantieniveau. Van deze acht modellen is er uiteindelijk één die voor beide zones geldt en die significante parameterwaarden heeft.

ii.3.3 Validatie

Na de regressie worden de drie potentiële modellen gevalideerd. Dit wordt gedaan door de uitkomsten van de modellen uit te zetten tegen de waarden van de variabelen die niet in de modellen zitten en door de modellen te vergelijken met instapdata uit 2004.
Uit de validatie met de 2008-data blijkt dat het aantal instappers enigszins wordt onderschat wanneer het werkelijke aantal instappers groot is en overschat als het werkelijke aantal instappers klein is. Verder wordt geconcludeerd dat er geen relaties tussen de residuen (de verschillen tussen de modelschattingen en de werkelijke waarde van het aantal instappers per inwoners) en andere variabelen zijn.
Met de 2004-data is de onderschatting van het aantal instappers bij grote aantallen instappers nog groter. Dit hangt samen met het feit dat in veel gevallen de toename van de frequentie tussen 2004 en 2008 niet terug te zien is in toenemende reizigersaantallen. Dit komt waarschijnlijk deels doordat het aantal instappers in een bepaald gebied ook afhankelijk is van weersinvloeden. Er wordt namelijk in beperkte perioden gemeten waardoor afwijkingen in het weer in die perioden duidelijk kunnen doorwerken in het aantal getelde passagiers. Deels komt het ook doordat er concurrerende lijnen zijn bijgekomen in een aantal zones. Verder is het mogelijk dat er een bepaald, zone afhankelijk, maximum aantal instappers is (een plafond). Als dit plafond is bereikt heeft het geen zin om de frequentie verder te verhogen. Tevens lijkt het zo te zijn dat het aantal arbeidsplaatsen per inwoner een beduidend minder belangrijke rol speelt in 2004 dan in 2008.

ii.4 Het model

Uit de analyse is het volgende model (tabel ii.1) naar voren gekomen. Dit model is geschat op basis van 20 van de 23 mogelijke zones. Het model is niet geldig voor zones waarin een treinstation is gelegen, de zone van de universiteit en zones met enkel bedrijventerreinen. De verklaarde variantie van het basisjaar van dit model is ongeveer 80%. Voor het controlejaar is dit ongeveer 64%.

R^2 = 0,804	Ongestandaardiseerde Parameters		Gestandaardiseerde Parameters	t	Sig.
	B	Std. Fout	Beta	B	Std. Fout
Constante	-0,117	0,036		-3,230	0,006
% inw. 15 t/m 24 jaar	0,00161	0,00133	0,183	1,212	0,246
Adresdichtheid	-2,46E-05	6,70E-06	-0,593	-3,669	0,003
Frequentie	0,00099	0,00013	0,852	7,869	0,000
Arbeidsplaatsen per inwoner	0,0229	0,0134	0,188	1,716	0,108
Punctualiteit	0,00185	0,00054	0,411	3,441	0,004

Tabel ii.1 Parameters van het model

In de tabel zijn de parameterwaarden weergegeven, de standaardfout van deze waarde, de gestandaardiseerde parameterwaarde, die wat zegt over de relatieve belangrijkheid van een variabele, en de t-waarde en het bijbehorende significantie niveau van de parameterwaarde.

De waarden van de verschillende variabelen, de consequenties voor het aantal instappers per inwoner en de bron van de variabelen wordt nu nader besproken evenals de relatieve belangrijkheid.

ii.4.1 Constante

De waarde van de constante is -0,117. Dit betekent dat wanneer de andere variabelen allemaal gelijk zijn aan nul, er een negatief aantal instappers per inwoner wordt voorspeld. Uiteraard kan het aantal instappers niet negatief zijn. Het geeft echter wel aan dat er bepaalde minimumwaarden nodig zijn voor onder andere de frequentie en de punctualiteit alvorens er gebruik wordt gemaakt van de bus.

ii.4.2 Percentage inwoners van 15 tot en met 24 jaar

Voor het percentage inwoners dat 15 tot en met 24 jaar is geldt een parameterwaarde van 0,00161. Een hoger percentage inwoners in deze leeftijdscategorie, een maat voor het aantal studenten, heeft dus als gevolg dat er meer instappers per inwoner zijn. Neemt dit percentage bijvoorbeeld toe van 10% naar 15% dan neemt het aantal instappers per inwoner toe met 0,008. Voor een zone met 4000 inwoners betekent dit een toename van het aantal instappers van 32.
De data komen in dit geval van de Wijk- en Buurtkaart, die elk jaar door het CBS wordt gepubliceerd. Deze kaart is alleen te openen met GIS-software. Op een meer geaggregeerd niveau (Postcode-4 niveau) zijn de data echter ook te verkrijgen via de site van het CBS. Voor Enschede is de data ook verkrijgbaar via de Buurtmonitor van de gemeente Enschede[ii].

ii.4.3 Adressendichtheid

De parameterwaarde van de adressendichtheid is $-2{,}46 * 10^{-5}$. Deze waarde is weliswaar veel lager dan die voor het percentage inwoners van 15 tot en met 24, maar doordat de waarden voor de adressendichtheid veel hoger zijn, is deze variabele relatief belangrijker. Wanneer de adressendichtheid toeneemt met 1000, vermindert het aantal instappers per inwoner met 0,0246. Voor een zone met 4000 inwoners komt dit overeen met ongeveer 98 instappers minder. Het teken voor de parameter is, in tegenstelling tot de verwachtingen, negatief. Dit is te verklaren doordat de bus in de dichter bebouwde gebieden rond het centrum van Enschede minder interessant is. De reden daarvoor is dat vanuit het centrum mensen en banen gemakkelijk met de fiets, de trein of te voet te bereiken zijn, terwijl er in de buitenwijken en omliggende gemeenten minder alternatieven zijn voor het busvervoer.
De adressendichtheid komt ook uit de Wijk- en Buurtkaart van het CBS. De (omgevings)adressendichtheid (OAD genoemd) van een bepaald adres is het aantal adressen in een straal van één kilometer van dat adres. Voor een buurt/zone wordt de OAD van alle adressen gemiddeld.

ii.4.4 Frequentie

De parameterwaarde voor de frequentie (aantal ritten per werkdag richting het Centraal Station van Enschede) is 0,00099. Een toename van één rit per uur (tussen 6:00 en 24:00 uur) leidt dan ook tot een toename van het aantal instappers per inwoner van 0,0178. Voor een zone van 4000 inwoners betekent dat een toename van 71 instappers.
De frequentie van een lijn is terug te vinden in de busboekjes van de vervoerder (nu: Connexxion).

ii.4.5 Arbeidsplaatsen per inwoner

De parameter voor het aantal arbeidsplaatsen heeft een waarde van 0,0229. Een toename van het aantal arbeidsplaatsen per inwoner van 0,2 naar 0,3 leidt dan ook tot een toename van het aantal instappers per inwoner van 0,00229. Voor een zone van 4000 inwoners betekent dit een toename van 9 instappers.
Het aantal arbeidsplaatsen in een zone komt, net zoals het aantal inwoners, uit het regionale verkeersmodel. Doordat dit model is ontwikkeld in samenwerking met de Twentse gemeenten is de informatie in het model erg betrouwbaar.

[ii] Te bereiken via: http://buurtmonitor.enschede.nl

ii.4.6 Punctualiteit

De parameterwaarde voor de punctualiteit is 0,00185. Een toename van de punctualiteit van 10 procentpunt leidt tot een toename van het aantal instappers per inwoner van 0,0185. Voor een zone met 4000 inwoners zijn dat 74 instappers.
De punctualiteit komt uit het Sabimos systeem, dat wordt gebruikt om bussen voorrang te geven en om te monitoren of bussen op tijd rijden. In de bijbehorende SabiMIS rapportage staat voor elke lijn weergegeven welk percentage van de bussen hoeveel te laat of te vroeg was in een bepaalde periode. Wanneer de punctualiteit vooraf moet worden geschat dan geldt dat de punctualiteit hoger is wanneer er vrijliggende busbanen zijn aangelegd en voorrangsregelingen zijn bij verkeerslichten. De punctualiteit ligt op dit moment grofweg tussen de 55% en 80%, met een gemiddelde van 65%.

ii.4.7 Relatieve belangrijkheid

In de vierde kolom van tabel ii.1 is te zien welke variabelen relatief het belangrijkste zijn. Wanneer de waarde van Beta groter is dan is de belangrijkheid ook groter, waarbij geldt dat een dubbele waarde van Beta betekent dat de variabele dubbel zo belangrijk is. Er kan dus worden gezegd dat de frequentie veruit de belangrijkste verklarende variabele is voor de vraag naar busvervoer. Daarna volgen achtereenvolgens de adressendichtheid (ca 70% van de frequentie), de punctualiteit (50%), en het aantal arbeidsplaatsen en het percentage studenten (ieder 20%).

ii.5 Aanbevelingen

Met de bovengenoemde uitkomsten van dit onderzoek kan een aantal vragen omtrent de drijvende krachten achter de vraag naar busvervoer worden beantwoord. Er zijn echter ook nog verbeterpunten te bedenken en mogelijkheden voor verder onderzoek aan te dragen, namelijk:

- OV-chipkaartgegevens gebruiken: Wanneer er betrouwbare en alle reizigers omvattende OV-chipkaartgegevens beschikbaar zijn, kan het onderzoek op een meer gedetailleerd niveau (met de haltes als onderzoeksobjecten) worden herhaald. Hierdoor zijn er meer onderzoeksobjecten en kunnen de lokale verschillen beter worden weergegeven, hetgeen mogelijk leidt tot een model met een hogere kwaliteit.
- Data van meerdere jaren gebruiken: Het model is geschat met behulp van data over één jaar (2008). Hierdoor is het aantal onderzoeksobjecten beperkt tot maximaal 23 en is het mogelijk dat het weer de tellingen beïnvloedt. Door meerdere jaren samen te nemen ontstaat er een model dat ook over de jaren heen geldig is en dat nog betrouwbaarder is aangezien er meer onderzoeksobjecten zijn gebruikt voor de schatting van het model.
- Busbereikbaarheidsmaat ontwikkelen: Er is gebleken dat geen van de maten van de bereikbaarheid van bestemmingen een significante en te verklaren relatie had met het aantal instappers per inwoner. Hierdoor zit er geen maat van de relatieve aantrekkelijkheid van de bus ten opzichte van andere vervoerwijzen in het model. Dit is echter wel wenselijk, wat betekent dat er een (andere) maat voor de busbereikbaarheid vanuit een bepaalde zone moet komen. Mogelijke maten zijn: de relatieve bereikbaarheid van arbeidsplaatsen/inwoners per bus (dus niet per openbaar vervoer als geheel) en de relatieve reistijd naar belangrijke bestemmingen.
- Centrum meenemen: Er kan verder onderzoek worden gedaan naar de manier waarop de centrumzone kan worden meegenomen in een model. Het aantal instappers is daar extreem veel hoger dan in andere zones aangezien veel terugreizen naar de wijken daar beginnen.
- Andere variabelen meenemen: In het onderzoeksontwerp is de keuze gemaakt om een aantal variabelen mee te nemen. Dit is echter een beperkt aantal en er zijn ook andere variabelen te bedenken die misschien iets van de verschillen in de vraag naar busvervoer verklaren. Er kan worden gedacht aan: variabelen met betrekking tot grondgebruik (waarbij een onderscheid wordt gemaakt tussen bedrijventerreinen, winkelgebieden en woongebieden), variabelen met betrekking tot verplaatsingsmotief (waarin onderscheid wordt gemaakt tussen woon-werk/school, winkel, ontspannings- en sociale ritten), persoonlijke voorkeuren en huishoudensamenstelling.

iii Preface

In this report I present the Master thesis of my Master Civil Engineering & Management that I enjoyed at Twente University in Enschede, The Netherlands. The research was conducted during an internship at Regio Twente in Enschede. Regio Twente is a public cooperation between 14 municipalities in the east of the Netherlands. Their Mobility department (Domain Living Environment) tries to keep the region accessible and is currently busy with the tendering of the regional public transport concession. During the first talks (in December 2009) with employees of this department it turned out that different questions regarding the traffic and transport were present of which the most urgent question was related to the demand for bus transport. All kinds of ideas existed on what variables were influencing this demand, but no clear, scientific, proof of these ideas was present. As I'm interested in public transport I accepted the challenge to come up with an answer to the question what the most important variables are that determine the demand for bus transport.
Early February 2010 I started writing a research proposal in which a literature study was elaborated to an outline and approach for the research that is described in this report. In the end the objective of the research was to come up with a quantitative model that can be used to determine a reasonably reliable estimation of the number of passengers on a local scale. Ideally a limited number of easy to obtain variables should be pinpointed that explain the largest part, say 80%, of the differences in passenger numbers.
After the proposal was approved I started the real research that included intensive data collection and management, data analysis, correlation analysis, regression and validation. This report was written during the different phases that were part of the research and at the end of my internship.

From time to time I could also do some small computation or calculation jobs. This included making (Google) maps, calculating costs of alternative bus schedules, error analysis in reports and making OD-matrices. In this way I was able to help making things clear or visualised and I also had some variety in my work, which I really appreciated. Another positive effect was that this improves the relations and contacts with the colleagues.

iv Dankwoord

Aan de totstandkoming van dit rapport en het onderzoek dat daaraan vooraf is gegaan ben ik bijgestaan en begeleid door een aantal personen.
Als eerst wil ik de leden van mijn afstudeercommissie bedanken. Mijn Regio Twente begeleider, Marco Berloth, voor het meedenken over het onderzoek en de feedback op de verschillende versies van mijn (tussen) rapportages. Ook heb ik het erg gewaardeerd dat hij mij, ondanks zijn drukke agenda, door verschillende trainingen heeft geholpen mijn presentatievaardigheden te verbeteren.
Ook wil ik mijn begeleiders van de UT bedanken. Karst Geurs vanwege de opbouwende kritiek en de nuttige verbetervoorstellen op de conceptversies en Tom Thomas voor zijn hulp bij het bepalen dan de onderzoeksmethodologie en de hulp bij het duiden van uitkomsten.
Verder wil ik graag mijn collega's van Regio Twente bedanken voor de geslaagde stageperiode van zes maanden. Ik heb de manier waarop jullie me opnamen in de organisatie als zeer prettig ervaren. Ook heb ik genoten van de reacties op de WK-Poule en de gesprekken tijdens de lunch. Speciale dank gaat uit naar Kim Wolterink en Gerda Dekker, die me welkom lieten voelen in hun (kleine) kamer waar ik tijdens de stageperiode mijn eigen plek had.
Frans van den Bosch van het ITC en Rogier van der Honing van Goudappel Coffeng hebben mij zeer geholpen doorat ze het mogelijk maakten dat ik ArcGIS en OmniTRANS kon gebruiken voor de bepaling van twee variabelen.
Als laatste wil ik graag mijn familie bedanken omdat zij mij altijd hebben ondersteund en omdat ze het mogelijk maakten om mijn stageperiode te combineren met mijn tijdrovende sportleven. Mijn vader wil ik ook graag bedanken voor het meedenken met het onderzoek en de feedback op de verschillende rapportages.

Enschede, augustus 2010

Erik Klok

1 Introduction

In the Dutch region of Twente, which lies in the east of the country, bus travel is offered with the name "Twents" (figure 1.1). Since 2006 a high rise in bus travel has occurred because of the introduction of this branding, cheap tickets and an increase of service (higher frequency and new busses)[1]. For the next concession period (from December 2013), the concession granting authority (Regio Twente) is interested how the current service can be increased and made more profitable. To do this it is important to know which types of services should be offered in what areas. Some lines that now have a high frequency and capacity might be better off with smaller busses when you look at their profitability. There can also be areas that have a high potential for bus transport but are not serviced in the current concession. To give better insights in the (potential) demand for bus transport it is important to know what transport infrastructure-, spatial- or individual characteristics determine the demand for bus transport. In this research the relations between the characteristics and the demand are described and a model is made that estimates passenger numbers given certain characteristics.

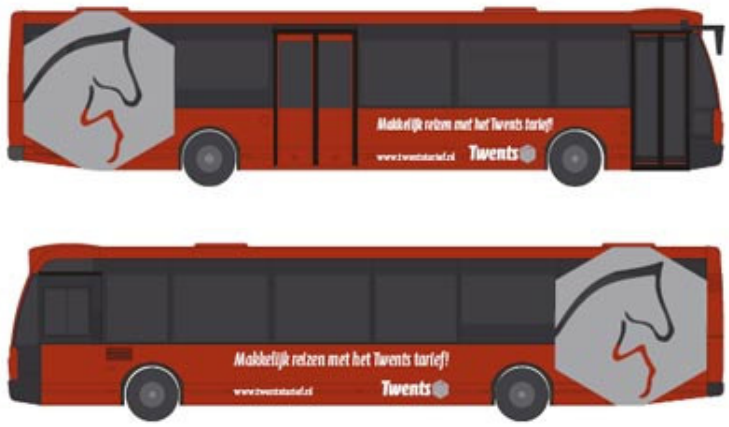


Figure 1.1 "Twents" bus design

Regio Twente

Regio Twente is a public cooperation of 14 municipalities in the eastern part of the Province of Overijssel in the Netherlands (figure 1.2). In the area about 620.000 people live. The policy domains that Regio Twente works at are: traffic & transport, economic affairs, safety, public health and tourism. Regio Twente does several things with regard to public bus transport. For instance, it is the concession granting authority in the area of Twente. It also makes a contribution to the exploitation of the public transport service with money that mainly comes from the national government. Regio Twente also determines the fare prices and controls the quality of the exploitation of the services.

Figure 1.2 Regio Twente in the Netherlands

In the current concession, Regio Twente pays the concession holder a price to exploit the bus services in the region, which does not cover the costs of exploiting the services. To overcome the difference between the costs and the regional contribution, the concession holder gets the revenues from tickets people buy. Higher passenger numbers hence are initially beneficial to the transport company that performs the bus services. If many people travel by bus however Regio Twente also benefits because less people are travelling with other modes and congestion might eventually decrease and road safety might increase (Regio Twente also deals with these policy domains).

Report outline

In this report the research is described from the problem description to the conclusion. First the problem is described in more detail research questions are posed (chapter 2). After that a literature

[1] http://www.keypointconsultancy.nl/PDF/20061211Busvervoer%20Twente%20groeit.pdf

study is described that gives insight in possible determinants of the demand for bus transport and possible methods of analysis (chapter 3). With the outcomes of the literature study the research methodology is determined, which is described in chapter 4. In the next two chapters the research area (chapter 5) and the subdivision of the research area into research objects (chapter 6) is described. After that the data used are described in chapter 7. Next the analysis steps, the correlation between variables (chapter 8), the regression models (chapter 9) and the validation of the models (chapter 10), are described. In the final chapter (11) a conclusion is drawn and the designed model is described.

2 Problem description

In the problem description the research background, -objective and -questions are introduced.

2.1 Research background

To have a good functioning transportation network it is important that demand and supply for mobility are aligned properly. For some transport modes, such as car or bicycle, this alignment is not quite an issue. When people that are willing to travel with a car or bike they have such a vehicle available at their house, at least when they bought such a vehicle. The only problems occur when demand is higher than supply which leads to congestion. Other (public) modes are however not available at any place and cannot move people from all possible origins to all possible destinations directly. Therefore certain stations and stops are built that are servicing a certain area. Between the stops and stations, lines are exploited over which people are transported. To determine how many and which vehicles per line are optimal it should be known what the demand is that one can expect in the catchment area of a certain stop or station. Usually this demand is however not studied in detail but only roughly estimated. Subsequently bus schedules are adjusted based on the observed patronage numbers. A more detailed analysis before a line is changed or added can lead to pilot projects with higher success rates.

On a more aggregate level in the planning of a bus network the demand should even determine where to place stops and how to assign lines to stops. The bus planning process is a very complex process that contains five activities: Network Design, Setting Frequencies, Timetable Development, Bus Scheduling and Driver Scheduling (Ceder & Wilson, 1986). Although the process is iterative over long time horizons the sequence of the process is as stated above. One of the most important inputs of the first step (Network Design) is the demand for bus travel in the area where the network is designed for. One of the problems however is that the demand is largely determined by the frequencies of the services (see: (Holmgren, 2007 & Paulley, et al., 2006), which is an outcome of the second step. It is therefore important to have a model that can determine a-priori how large the demand will be, given certain characteristics.

Besides frequency of services there are many more variables that influence the demand for public transport. Examples of them are: price of tickets, people's income level, car ownership, quality of service, trip purpose, price of substitutes and level of urbanisation (Holmgren, 2007, Bresson, Dargay, Madre, & Pirotte, 2003, Paulley, et al., 2006 & Souche, 2010). These variables have all been studied many times. Most of these studies use averages over large areas to explain variations in bus use. Another characteristic of these studies is that data from different years are used to determine for instance the relation between variables such as car ownership or income and the demand for busses. Variables that have not yet been studied intensively, but might have an impact on the demand for bus travel are: educational level, bicycle ownership and many others.

In order to determine how the bus system in its concession area is functioning and whether or not improvements can be made, Regio Twente is interested in a model, or an extension to the car traffic model they currently use to determine the demand for bus transport. When the demand can be determined, Regio Twente can check whether some lines should be changed or whether it would be beneficial if new stops or lines are added. In this way the network should become more efficient, which means that the busses transport as much people as possible given their capacity. The outcomes could mean that on some lines smaller or fewer busses should be deployed, other lines may be terminated completely and some neighbourhoods or villages should get bus lines connecting them with other villages and neighbourhoods. Because the information that you need is at a small scale (neighbourhood/bus stop level), it is important to know how differences in the previously mentioned possible determinants

Figure 2.1 Concession areas in the Netherlands

between small areas (for instance neighbourhoods or districts) relate to the demand for public bus transport in those areas. Current literature however mostly studies average variable values over whole cities or even counties (like Bresson et al, 2003). The usability of the outcomes of those studies is therefore probably not very high. Another important implication of this purpose is that you have to look at differences between zones in a certain year rather than on differences between years in a certain zone, which is done in most studies mentioned earlier.

The concession area of Regio Twente lies in the east of the Netherlands in the province of Overijssel (area 12 and part of 13 in figure 2.1). In the area there are three cities (Enschede, Hengelo and Almelo) that have urban bus services that are also part of the concession. The other lines are regional (rural) services that run mainly between villages and cities. The current concession holder is Connexxion, whose concession period ends in 2013.

2.2 Research objective

It is clear that Regio Twente wants to have a model that gives insight in the demand for bus transport in their concession area on a local scale. Therefore more knowledge is needed of the differences in demand for bus transport on a high spatial resolution (neighbourhood/bus stop level). The demand should be dependent on different characteristics of the transport system, living/working areas and people living in the areas. Because taking the whole concession area into consideration would be too complex and time consuming for the nature of this research, the research objective is:
Develop a model that describes how transport infrastructure-, spatial- and individual characteristics determine the demand for bus transport in the city of Enschede and the rural municipalities of Haaksbergen and Losser.

2.3 Research questions

From the research objective the following main research question can be formulated:
How can the relations between the demand for bus transport and transport infrastructure-, spatial- and individual characteristics in Enschede, Haaksbergen & Losser be modelled?

This question can be subdivided into several sub-questions that are used to answer the main research question. The sub-questions are:

1. What determines the demand for bus transport? (Theory)
2. What are the most important variables that need to be included in the bus demand model?
3. How can the study area be subdivided into research objects?
4. What data are needed?
5. What are the (quantitative) relations?
6. How can the model be validated?

These questions are translated into the following research description.
First an overview of variables that possibly influence the demand for bus transport is given. This overview gives an answer on the first sub-question (*What determines the demand for bus transport?*). Based on the theory, a short description of the currently used car traffic model and a description of methods that are used to model relations between variables it is described what variables to study and how they are studied. Doing so gives an answer on the second sub-question (*What are the most important variables that need to be included in the bus demand model?*).
After that the research area and its subdivision into research objects is described more detailed (sub-question 3: *How can the study area be subdivided into research objects?*). After that an overview of the data that are needed to study the relations is given. In this step also the collection, storage and management of the data are described. This is all related to the fourth sub-question (*What data are needed?*).
When the data are combined and correlation is checked it is possible to make a model of the relations between the different variables with the statistical tool SPSS. Based on some requirements on the model parameters then some potential models are highlighted which answers the fifth sub-question (*What are the (quantitative) relations?*).
When the first relations are determined it is important to validate the model. Off course it is checked whether the outcomes are in line with available literature. The increase in bus patronage that was mentioned in the introduction is an extra opportunity to be used in a validation step. The model that is the product of the previous steps therefore is tested on its ability to explain the rise in the number of passengers. It is also tested whether the models wrongly do not take some variables into account. With the outcomes an answer is given to the last research sub-question (*Can the model be validated?*).

3 Literature study

In the literature study an overview of possible determinants and methods to model modal choice behaviour and relations between variables is given.

3.1 Possible determinants

There are many variables that can be related to the demand for bus transport. Those determinants can be classified in different ways. An important difference between variables is that some describe a characteristic of a built environment, while others are about individual or household (socio-economic) characteristics (Van Acker & Witlox, 2010). Within the built environment the determinants can again be divided into characteristics that relate to the origin of a trip while others relate to the destination of a trip. In between the origin and the destination, the network facilitates travel. Because they are different for different origin-destination relations, these characteristics are also taken into account for this research. A last category of possible determinants is the category of personal preferences. In figure 3.1 a schematic overview of the different categories of variables is given.

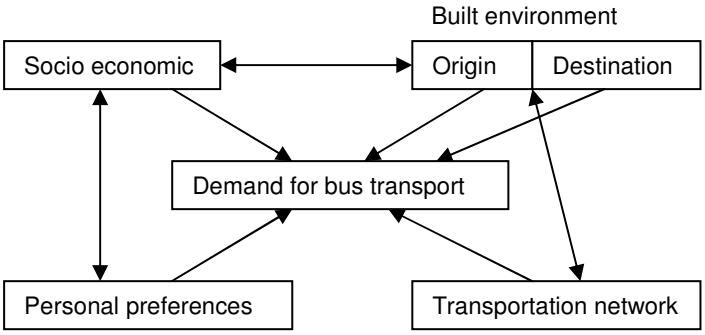


Figure 3.1 Variable categories that relate to the demand for bus transport

In the figure it can be seen that the different categories of variables do not only relate to the demand for bus transport but also to each other. The built environment of the origin of a trip for instance is related to the socio economic characteristics of the people that live in that origin. Also relations between personal preferences and socio economic characteristics and between the transportation network and the built environment are present.
The variables in the different categories are now described in detail. In appendix A for each category an overview is given of the different variables that can be related to the demand for bus transport and variables that are related to those determinants. Also relations between possible determinants are included in the overview.

3.1.1 Socio economic

Socio economic characteristics are related to individuals and households. Examples of socio economic characteristics that have influence on the demand for bus transport are whether or not a car is owned, the income of a household and the composition of a household. These (and other) characteristics and their influence on the demand for bus transport are described here (see also figure A.1 in appendix A).

Car ownership
Car ownership is often indicated as one of the main determinants for car use (Van Acker & Witlox, 2010). When people have a car they in most cases also use their car, or, at least, they use it more often than people that do not have a car. Because (relatively) more car trips are undertaken the number of bus trips in general decreases. The relation between car ownership and the demand for bus transport is hence often found to be negative. Paulley et al. (2006) for instance report that people that own a car have a demand for bus transport that is more elastic than people who do not own a car. Holmgren (2007) even showed that there is a direct negative elasticity between the demand for bus transport and car ownership. This means that an increase in car ownership leads to a decrease in the demand for bus transport.
Car ownership itself also has some determinants. Van Acker & Witlox (2010) conclude that the two most important determinants for car ownership are income and owning a driving licence. When people have a high income it is relatively easy for them to buy a car and therefore they also own, on average,

more cars than people with a low income. The relation between owning a driving license and car ownership is also very clear-cut: Because people that do not have a driving license are not allowed to drive a car owning one is useless. Other variables that also are related to car ownership are: gender, age, marital status, educational level, employment status and household size. Educational level, employment status and income are however highly related to each other and have comparable relations to car ownership (Van Acker & Witlox, 2010).
When people "own" a business car they are strongly encouraged to use the car as mode for transport (Maat & Timmermans, 2009). This is mainly because people with a business car do not pay for the use of the car, at least not for travelling to, from and for their work. Alternative modes are therefore always more expensive than using the company car for work related travel and in some cases also for travel with other purposes. Car ownership is however also very much correlated with the income of people while high income people are more likely to have a company car. Maat & Timmermans (2009) even concluded that when company car ownership was taken as a determinant for car use, a significant relation between income and car use was absent.

Bicycle ownership
The car is not the only substitute for the bus. People can for instance also choose to make a trip by bicycle instead of taking a bus. Owning a bicycle is a prerequisite for this to be an option.
In many researches bicycle ownership is not taken into account when determining elasticities or effects on the demand for bus transport. The only research that mentions cycling clearly is (Fitzroy & Smith, 1998). Interesting to see is that in the period that public transport (PT) ridership increased also the modal split for cycling increased, which could mean that both modes are not substitutes but complements. In the same period the modal split for walking reduced very much. Whether the increase in bicycle use and PT use are related, or whether bicycle use increased just because of the decline in walking, therefore cannot be ascertained. One would however expect that when there are more bicycles in a certain area people use the bus less. The only way that the bicycling can be complementary to bus use is when people use it for access or egress transport.
Bicycle ownership is very common in the Netherlands and therefore bicycle use probably does not differ that much over people with different socio-economic characteristics. Probably the only interesting characteristic that influences bicycle ownership is age because older people own less bicycles and young children first need to learn how to cycle and also own less bicycles. Because the bicycle is not seen as a poor man's mode in the Netherlands socio-economic characteristics like education and income probably do not really influence the number of bicycles owned. Only people that are not able to cycle will own fewer bicycles and also use them less.

Income
Another important determinant of the demand for bus travel is the income of people. Bus transport is often classified as an inferior good, meaning that an increase in income leads to a decline in the demand for the good/service (Souche, 2010 & Holmgren, 2007). This is mainly because people that have more money to spend are using the car more for their transportation. There is however also research that shows that an increase in income levels does not necessarily have to lead to a smaller demand for bus transport (Bresson et al, 2003 & Fitzroy & Smith, 1998). A reason for these contradicting findings can be that all the studies use time series to determine the relations between income and demand for bus transport. In some cases the demand went up and in others the demand went down, while the income increased in all of the cases (as in all western countries since World War two). The reason for the increase of the demand for bus transport in one case and decrease in other cases is probably attributable to other variables than just income. Income probably is an important determinant for differences in the demand of bus transport on a local scale. There are however also studies that suggest that income does not really influence the demand for transport, but only car ownership. The differences in car ownership then are the main determinants of differences in the demand for bus transport (Van Acker & Witlox, 2010).
What is also interesting is that people that are really poor maybe cannot afford to travel by bus. They might cycle more because using a bicycle is free. If this really is the case the relation between income and the demand for bus transport is not linear but polynomial.
Income is highly correlated to educational level and employment status (Van Acker & Witlox, 2010). These two variables hence are also determinants of the demand for public transport, but also of the income of people. Because similar relations are present, studying all three variables is unnecessary. Income also is related to gender and age. Younger people for instance have lower incomes than older people and males usually get more money for the same functions than women (Erdem & Houben, 2008).

Educational level
The income of people is highly related to the educational level they have enjoyed. There is however special Dutch policy that might lead to different relations between educational level and the demand for public transport in the Netherlands. People that have enjoyed high education (since 1991) have had a free travel Public Transport-card during their time as a student. One of the ideas behind this card was that students that get used to travelling by public transport are going to use public transport later on in their life. When however the bus network you used as a student was very poor this might also work counterproductive.
Because, compared to the total population, only a limited number of people have had such a card the influence of the educational level (next to income level) probably is not very large.

Driving license
To use a car by yourself it is obliged to have a driving license. People without a driving license can hence (theoretically) only use a car as a passenger. Their number of car trips is likely to be relatively lower. The other modes therefore take a larger share of the trips of people without a driving licence than people that do have a driving license. The share of bus transport is therefore also probably higher for people that do not have a driving license.

Household composition
Different characteristics of the household composition can also be determinants of the demand for bus transport. Household size for instance influences the travel distance in such a way that larger households travel further. Larger households are also more dependent on car use (Van Acker & Witlox, 2010), which means that they probably use the bus less. An advantage of the car as a mode of transport is that using it costs the same for 5 persons as for 1 person. For bus use however travelling with 5 persons is 5 times as expensive as travelling alone.
It is also important how many people work in a household. In dual-earner households travel distances tend to be higher. Also car dependency, and use, are higher in dual-earner households. The study of Maat & Timmerhuis (2009) shows however also that in dual earner households with more than one car, cars are not used for commuting in many cases (about 40%).
When young children are in a (dual-earner) household men tend to leave the car at home more. This is because women in most cases are responsible for the transport of the children to and from school. Women hence profit most from the car and men commute with another mode (Maat & Timmerhuis, 2009).

Other
Other characteristics that can have a relation with the demand for bus transport are being a student or disabled.
Students get a PT-card with which they can travel for free during week or weekend days. Because using the bus is free for them it can be expected that they use the bus more often (Fitzroy & Smith, 1998). The number of students in a certain zone therefore probably is an important determinant of the demand for bus transport.
For disabled persons multiple considerations can be made. It is hard for them to drive, so car use will not be very high. It is possible to get discount tickets for the bus so they can use a bus more often than other people, but also special transport means are deployed for them. Using other modes (e.g. Regiotaxi) reduces the number of bus trips from disabled people.

Indirect determinants
There are also variables that only indirectly influence the demand for bus transport such as someone's age, gender and culture and whether someone is a house owner or not.

Age
Age does also matter for the demand for bus transport. Young people (under the age of 18) cannot own a driving license and will therefore use a car less and might use busses more often. Car driving becomes more difficult when people get older which leads to lower car use of elderly people. Because they can also travel by bus with a discount their relative bus ridership can be expected to be higher than people in other age categories.
Children under the age of 11 also get discount. The number of people in that age category can also influence the demand for bus transport. Children and elderly people together in some areas in

England produce even 40% of all bus trips (Rye & Scotney, 2004). It should be studied what age groups should be included in a model for the demand for bus transport.

Gender
It has been shown that males have higher elasticities for bus use than females (Paulley, et al., 2006). The fact that women travel with public transport more often (Van Acker & Witlox, 2010) seems to support this conclusion. The reason for it can partially be that males are more likely to own a car. Women also travel over shorter distances, which can mean that they also use bicycles more often.

Culture
Also culture, or being part of an ethnic minority, can play a role in the demand for bus transport. Research has shown that about 75% of Britain's whites uses a car for commuting while this is only about 55% for ethnic minorities (Gautier & Zenou, 2010). As a consequence a larger portion of ethnic minorities depend on public transport (33% to 14%). Gautier & Zenou (2010) conclude that most of these differences are related to differences in income. If that is the case culture is not a good variable to describe differences in the demand for bus transport next to income. The question however is whether the differences in Britain (and also the US) also appear in the Netherlands. Besides income it can also be that people of different ethnic minorities have more or less affinity with bus transport leading to more or less bus use.

House owner
Whether somebody owns a house or rents it is another indirect determinant. Renters are disproportionally poor, young, located in denser multifamily housing that may lack parking (Kuby, Barranda & Upchurch 2004). It hence is related to many other variables that possibly have impact on the demand for bus transport.

3.1.2 Built environment

The built environment is related to zones/neighbourhoods between which travel takes place. Because travel takes place between an origin and a destination, the different determinants are divided between characteristics related to the origin and characteristics related to the destination of a trip (see also figure A.2 in appendix A).

Origin
The design of an origin zone or neighbourhood determines to a large extend how well accessible it is for different modes and also how many inhabitants or jobs are serviced by a bus system. The four categories of built environment characteristics that are defined by Van Acker & Witlox (2010) are used to describe the relations between the built environment of the origin and the demand for bus transport.

Spatial design
The design of a neighbourhood is crucial for good public (bus) transport possibilities. There are several variables that can be influenced with spatial design. They are: the types of houses, the spatial spread and the number of parking places.
There are many types of houses ranging from free-standing houses to apartments. The individual houses in those categories can again be small (cheap) or large (expensive). Choices that are made about the houses hence have huge influences on the people that are going to live in a neighbourhood. There hence are clear relations with the socio-economic characteristics of the inhabitants.
The spread of buildings is also important for the possibilities of bus transport in an area. The houses and other buildings can be concentrated in small areas or spread in large areas. When the former is done it is easier to serve all houses and less stops are needed which leads to smaller travel times. The resulting network is more efficient and more competitive to other modes. The demand for bus transport hence is higher (Van Acker & Witlox, 2010).
Another important design issues is the number of parking places. The more parking places there are in a living area, the better the area is accessible for cars and the more interesting car use becomes (Van Acker & Witlox, 2010). Accompanied with the number of parking places is the design decision whether or not people have to pay to park. When this is the case car ownership and also car use should be lower then when parking is free.

Population density
People in rural areas with low densities use the car more than people living in densely built urban areas (Souche, 2010 & Van Acker & Witlox, 2010). This is because bus transport in most of the rural

areas is not a good alternative. Elasticities are therefore higher in rural than in urban areas (Paulley, et al., 2006). It is also showed that people living in more densely populated areas travel more kilometres and undertake more trips by bus weekly (Balcombe et al, 2004).

Spatial diversity
Not only the population density in a certain zone is important for the demand to bus trips, but also the density of jobs and services plays a role. Especially the combination of different functions (living, working, shopping, leisure) is important, which is often referred to as spatial diversity. A more diverse neighbourhood usually produces more PT-trips. The average distance of the trips is also smaller which leads to more trips with slow modes such as cycling and walking (Van Acker & Witlox, 2010). One of the reasons for this is that in diverse areas always people are present. This leads to more social safety which is a real issue when travelling by public transport. People have to feel safe when they walk to a stop and when they have to wait for the bus. When streets are empty the safety feeling of people is low which leads to less bus use, especially in off-peak periods.

Accessibility
The accessibility of the neighbourhood by the different modes is also important in peoples modal choice. Accessibility in this context is ”the ability to reach activities or locations by means of a travel mode” (Geurs & van Wee, 2004). When the accessibility of a certain mode is higher (more locations or activities can be accessed with the mode in a reasonable time) people use the mode more (Van Acker & Witlox, 2010). For the different modes there are different reasonable travel times and distances. Walking for instance is only used for short trips (till 2.5 km), the bicycle for longer trips 0.5-7.5 km, the bus for trips between 1 and 30 kilometres and the car is used for trips larger than 1 kilometre (Rijkswaterstaat, 2008).

Destination
For the destination of a trip more or less the same sub-categories can be recognised. The trip purpose however is really related to the destination and not to the origin of a trip. Therefore the relation of the purpose of a trip to the demand of bus transport is described in this part of the report.

Spatial design
The spatial design of the origin destination is important for mode choice, the spatial design of the destination location is however also important. When people are considering different modes they might use for a trip they for instance take into account whether or not there are parking places at the destinations and whether or not parking is for free. The spatial design also influences egress distances of public transport. Because people in general only can walk as egress mode it is very important that destinations are closely located to bus stops. There are however possibilities to give people more alternatives as egress mode. At train stations in the Netherlands there are for instance “Public Transport bicycles” that people can use to reach their destination. In this way the accessibility by public transport is increased.
Also the location of services and jobs in relation to the origins is important. When jobs and services are located closely to origins it is likely that people travel by bicycle in between the two locations. When the distance becomes larger and fast bus lines are exploited in between the zones bus use is more likely while when bus accessibility is not good while car accessibility is very good people are likely to use the car for their trip.

Spatial density
Spatial density is also very important in the destination zone. Maat & Timmermans (2009) for instance refer to studies that state that the job density in a destination is most important for the mode people choose to commute. It is said to be even more important than the residential density in the origin zone. High densities of jobs are usually concentrated around PT-junctions or in central business districts. These locations are hence very good accessible by public transport, which explains the relatively greater modal split for train but also for bicycle (Maat & Timmermans, 2009). Whether this also accounts for bus trips is unknown.
Besides job density also service density can be important. When there is a large concentration of services the number of potential customers that comes to the area is higher which makes exploiting bus services more reasonable. People travelling to larger shopping areas (city centres or malls) therefore are more likely to use public (bus) transport.
What is also interesting is the number of college (and university) enrolments in a certain zone. These trips are different than other trips because most of the people that undertake these trips have a

discount card in the Netherlands. Therefore many trips are undertaken to these zones by bus (and train). Scholars that enjoy secondary education (like: MBO) are also less mobile than older people. Therefore also the number of secondary education enrolments in a zone probably is related to the number of boardings in that zone.

Trip purpose
The purpose of a trip also has influence on whether or not people are elastic to bus transport or not. On average trips for commuting to work or education are the least elastic, which means that effects of changes in price or service will influence them least. Leisure trips however are more elastic to the price and quality of services (Paulley, et al., 2006). The reason for this is probably that the number of alternative modes and destinations for leisure trips is higher. Also, because leisure trips are on average made less often, using (temporal) discount tickets for these trips is less beneficial than for daily trips to work or school.

3.1.3 Transportation network

The transportation network facilitates the travel from origin to destination. There are different characteristics that are possible determinants for the demand for bus transport. Not only characteristics of the bus network are described, but also network characteristics of other (substitute) modes are included (see also figure A.3 in appendix A).

Quality of service
The quality of service is an important determinant of the demand for bus transport. The better the quality of service the higher the demand for bus transport becomes. The quality of service is determined by a large number of variables. Most of them relate to the perceived quality by travellers and non-travellers. Those variables are: frequency, area covered, reliability, punctuality, seating probability, need for transfers, bus appeal, travel information, the presence of designated infrastructure and park or bike and ride facilities.
An important part of the quality of service is the frequency. A higher frequency leads for instance to less waiting time. When people can access bus lines with high frequencies their bus use is expected to be higher (Paulley, et al., 2006).
The percentage of the total area that is serviced by a bus network, often referred to as cover, also determines the quality of a bus network. When large percentages are covered more destinations can be reached which should lead to high numbers of bus patronage. The cover is related to the network design and the stop spacing. When the stops are located with large mutual distances the cover of the network is smaller than when close spacing is used. In figure 3.2 this is shown. Disadvantage of close spacing is however that the travel times become larger because busses have to stop more often (Kuby et al, 2004).

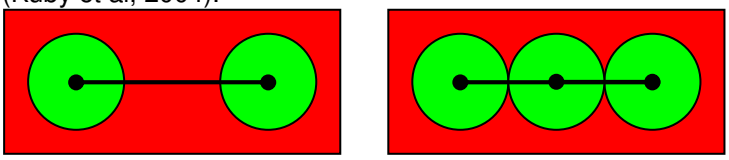


Figure 3.2: Distant (left) & close (right) stop spacing, green = covered area

Reliability is related to the amount of services that falls out. The more services fail on a certain line, the less reliable the line is. Related to the reliability is the punctuality (which is in fact travel time reliability). The punctuality is a characteristic of being able to perform a certain task (a bus trip) according to schedule. When all services are on time (or with a small delay) the service is very punctual. When the route of a bus line is heavily congested the travel times increase and as a consequence more services are late which leads to a lower punctuality. This problem can be solved by building designated infrastructure for busses. Busses can then use specific lanes or streets where no other traffic is present which makes the travel times more reliable. The presence of special bus lanes can also increase the comfort feeling of people which directly leads to a higher quality of service. Systems that, if needed, give busses priority at intersections can also increase the punctuality of the bus system.
Seating probability can also have a relation with the demand for bus transport of people. When the chance that you have a seat is higher people can perceive the trip as more comfortable than when they have to stand often. In general it can be said that people boarding at stops that are closer to the starting point of the line have a higher seating probability.

A quality aspect that is determined by the network design is the need for transfers. For travellers it would be ideal that all destinations can be reached from an origin without having to transfer. This is however impossible because of the cost constraints of the system. The number of transfers however should be minimised to make the experience of the travellers more positive. Transfers are heavily disliked by travellers. For busses a transfer is even valued as 21 minutes of in-vehicle time (Paulley et al, 2006).
Also bus appeal can play a role in the quality perception of travellers. When a bus is old and badly maintained it is likely that some people will not use it even if the travel time is much shorter than travelling by other modes. Otherwise it is also possible that busses that are new, nicely decorated and "green" attract travellers that might travel by other modes when a standard bus is used.
Which travel information is present at stops can also influence traveller's perceptions of the bus system. There are different kinds of information, that can be categorised as static (scheduled bus departure times, next stops, network maps) and dynamic information (actual departure times, delay). The more (actual) information is present the higher the quality perception of travellers.

Travel time
Travel time is another determinant of the demand for bus transport. Travel time by bus consists of different parts because the bus is not a door-to-door mode. Those parts are: in-vehicle time, access & egress time, waiting time and transfer time.
In-vehicle time determines a large part of the travel time. When people have to travel long times or with large detours the relative attractiveness of the bus becomes less. Related to in-vehicle time is the centrality of a stop in the network/line (Kuby et al, 2004). The centrality tells something about the travel time towards all other stops. Centrality can be reversely related to seating probability because when a stop is located more centrally the seating probability is usually lower.
Other important variables to keep in mind are the access and egress times that are needed to reach a bus stop or your desired destination (Paulley et al, 2006). When people have to travel longer before and after they use the bus it is less interesting to use the bus. There are different variables that influence the access and egress times, like the distance to a stop and the presence of park and ride or bike and ride facilities. Park and ride or bike and ride makes it possible to access with other modes than walking that are usually faster which makes that the same access distance can be travelled in shorter time. Those facilities usually are only located at origin stops of work and school trips.
Waiting time is also important in people's modal choice decisions. One minute waiting time is considered to be a larger disutility in travelling than in-vehicle time (Paulley et al, 2006). The same negative relationship can be found in the elasticity of the demand for bus transport with respect to waiting time (-0.64). Waiting time is influenced by the frequency of the services and the quality of service (punctuality, reliability). When the frequency is higher the average waiting times is lower because the follow up times of the vehicles are shorter. Less punctuality and reliability leads to extra waiting times at stops because vehicles arrive there later or not at all.
Not only the number of transfers but also the transfer times are important when the attractiveness of the bus to potential travellers is studied. Transfer times increase total travel time and hence reduce the relative attractiveness of the bus over other modes. Transfer times are also influenced by the frequency of services and, moreover, the frequency of the lines that is transferred from or to. When the destination line has a high frequency transfers are always short. When the frequency of the origin line and the destination line is low the arrival and departure times should be synchronised in order to make the transfers short while information should be given when the destination line has a low frequency and the origin line has a high frequency (Van Berkum, 2009).

Price of fares
Bus demand is not only depending on the quality of services and the travel time. The price of fares, for instance, also plays a role (Fitzroy & Smith, 1998) & (Souche, 2010). In general higher prices lead to lower patronage.
There are also special tickets that can be bought if you use bus transport regularly. There are for instance, weekly and monthly cards and also subscriptions for a whole year on certain lines or for certain combinations of zones. People that bought these tickets have less marginal costs for an extra bus trip and will, because of that, travel more.
A special group of travellers that do not have to pay for public transport are the students. They are already described in the socio economic characteristics of an origin zone and is not discussed in more detail here.

Substitutes
Not only the availability of substitutes (which is an origin based socio-economic variable) plays a major role in the demand for bus transport, also the price of these modes does. When it is more expensive to use a car the disutility of travelling with the car becomes higher which makes other modes (such as the bus) more attractive. In research the price of petrol or the overall costs of cars are used as a measure for the price of this substitute (Holmgren, 2007) & (Souche, 2010).
It is also important how accessible a zone is for the different substitutes. If a neighbourhood for instance is good accessible by car (because it is near a highway) people will use cars relatively more than in neighbourhoods that have a bad car-accessibility (Souche, 2010). Also the accessibility by bicycle is important, but in the Netherlands the accessibility by bicycle is quite good everywhere, so differences in bicycle accessibility may be small. For the competitiveness of the bus to other modes the travel times of all the modes are most important. These travel times depend on the distance travelled and disturbances on infrastructure that the substitutes make use of.
The distance between origin and destination also is important because large distances are often travelled by train and car while for small distances the bicycle is a good alternative for busses. The average distance to a destination hence can be another determinant of the demand for bus transport.

3.1.4 Personal preferences

Different people think differently on travelling. Some people like travelling by car because it gives them a safe or autonomous feeling, while others feel stressed when they drive a car. Those personal preferences (figure A.4 in appendix A) also influence the mode people are choosing for their trips. Gardner & Abraham (2007) mention the following factors of individual preference that influence the modal choice of people: stress, excitement, uncertainty, safety, enjoyment and feeling of autonomy. All the variables can be related to car, bus or other traffic modes. Some people feel stressed when they drive a car while others can perceive making bus transfers very stressful. Also people can be excited about car driving (car lovers) or about using public transport.
Something that is also influencing the modal choice of people is the fact that when using mass transport people are less busy during transportation (they are transported by a bus or train driver) which means that they can do something else themselves. People that have to read much or can use a laptop for their work hence can do some of their work in the bus which saves them time on their work. Other people however don't have these opportunities.

3.2 Travel demand modelling

Much research has been done on topics that relate to the modelling of transport behaviour and travel demand. Research for instance has been done about travel demand modelling and data analysis. These two topics are therefore be covered in this paragraph.

3.2.1 Four step model

Travel demand is usually modelled using a four-step model (Ortúzar & Willumsen, 2001). The four steps of the model are: trip generation, trip distribution, modal choice and route choice. Those steps are all described shortly. The Regional Traffic Model Twente (Friso & Van der Honing, 2006) also works with these steps.

Trip generation
Travel generation is the step where the number of trips from (production) and to a certain zone (attraction) are determined. In simple modelling these can only be dependent on the number of inhabitants or jobs in a certain zone. In more complex models also other characteristics of people in a zone can be taken into account, like car availability, income or size of household. Usually in morning peak periods there is a simple linear relationship between the number of inhabitants and the number of trips produced and also a linear relationship between the number of jobs and the trip attraction. In the evening peak production depends on the number of jobs and attraction on the number of inhabitants (Ortúzar & Willumsen, 2001).

Trip distribution
When the number of trips in a certain zone is known it is possible to determine from which origins to which destinations trips are made. This is determined using a gravity model that uses the resistance between different zones and the amount of trips from and to the zones to determine the number of

trips in between the zones. The basic idea behind this model is that when travel takes longer or is more expensive it is less likely that travel takes place. Between zones that are further away from each other less trips are performed than between zones that are closer to each other (given the same trip production and attraction numbers). Resistances between all zones are computed in this step and are used to make an Origin-Destination matrix that gives an overview of how many trips from each zone to all other zones are undertaken (Ortúzar & Willumsen, 2001).

Modal choice
For all the different Origin-Destination pairs it is then possible to determine how many people are going to take the different possible modes for a trip. This is usually done using logit models. Logit models use the disutility of travel between the zones for the different modes. Travelling is a disutility because it costs time and money. The larger the disutility of a certain mode compared to other modes the smaller the modal share of that mode. Factors that can be included in the utility function of a mode are: mode specific parameter (like/dislike of a certain mode), in-vehicle travel time, travel costs, access & egress travel time, waiting time and comfort. The basic assumption behind the utility function is that people want to maximise their utility and therefore choose the alternative that has the lowest disutility (Recker, 2001). Because not all people do so the logit model is used to translate differences in utility to percentages of trips made by the different modes (Ortúzar & Willumsen, 2001).

Route choice
When the modes for all the trips are known the trips can be assigned to the network. For most OD-pairs more than one route can be used. Multiple methods are known for the assignment of the trips to the different routes. Examples are: All-or nothing (all trips use the shortest route), stochastic methods (a proportion of trips to routes based on differences in travel (time) costs), congested assignment (Wardrop's equilibrium/incremental assignment/successive averages). For public transport assignment some extra problems can occur. This is because bus transport is not a door-to-door mode but needs to be accessed and egressed and because people travelling by public transport have to wait and transfer (Ortuzár & Willumsen, 2001). More about public transport in the paragraph (3.1) about possible determinants of the demand for bus transport.

3.2.2 Modelling relations

There are two methods that can be used to describe the relation between different variables and the demand for bus transport. These methods are regression analysis and discrete choice models and are described here. First however correlation analysis, which is used to check whether or not the value of one variable is related to that of another variable, is described. Last part of this paragraph introduces the concept of elasticity which is used to describe people's (travel) changing behaviour with respect to changes of variable values.

Correlation analysis
Correlation exists when different variables' values are clearly related to each other. When variables are correlated the value of one variable can be estimated with the value of the other variable with a high precision. In many cases the correlation can be explained because the two variables say something about the same phenomenon. When several measures for the same possible determinant are considered it is presumable that the variables correlate. In contrast, a lack of correlation arises questions on the credibility of at least one of the measures. In a correlation analysis it is checked which variables correlate and what the consequences of the correlation are.

The amount of correlation is usually expressed in the Pearson's correlation coefficient.

Pearson's Correlation coefficient = $\frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2}\sqrt{\sum (Y_i - \overline{Y})^2}}$

Where:
$\mathrm{cov}(X,Y)$ = Covariance of variables X and Y
σ_X = Standard deviation of variable X
X_i = Variable value of X of research object i
$\overline{X}$ = Sample mean of variable X

The correlation coefficient hence is the covariance of the two variables divided by the product of the standard deviations of the two variables. The covariance is the expected value of the product of the deviation of a single value from the mean. When for a certain research object the relative distance to the mean is equal for two variables the coefficient is close to 1. When the relative distance of one variable is equal to the negative relative distance of the other variable the coefficient is close to -1. When there is no relation at all between the variables the coefficient is close to 0. In figure 3.3 different examples of related data are shown with the corresponding correlation coefficients.

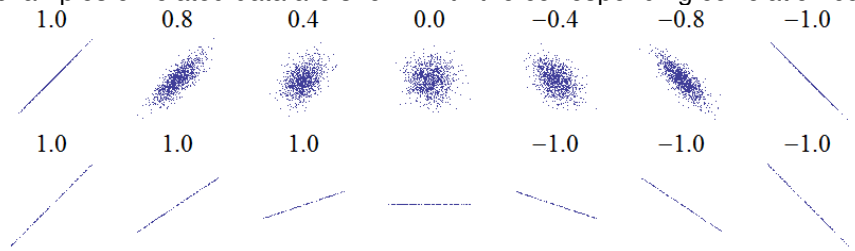


Figure 3.3 Correlation coefficients of different datasets

Regression analysis

Regression analysis is used to describe relations between dependent and independent variables. It helps to understand how the dependent variable changes when independent variables are varied. The result of a regression analysis is a function/model in which the value of the dependent variable (D) is described as a function of the independent variable(s) (I). A regression function can look like this:

$$D_i = \alpha_1 I_{1,i}^{\beta_1} + \alpha_2 I_{2,i}^{\beta_2} + \varepsilon_i \qquad (3.1)$$

In this function alfa (α) and beta (β) are model parameters and epsilon (ε) is the error term.

Types of regression

There are different types of regression analyses. The different possibilities are described here.

Simple/multiple

Both simple and multiple regression analyses can be performed. Simple regression analysis covers only one independent variable while a multiple regression analysis considers multiple variables at a time. Kuby et al. (2004) use a multiple regression because “it is flexible, widely used, easily understood, able to handle dummy variables and because the resulting model can be used for predictive purposes in other cities”. To check whether different independent variables of a multiple regression analysis are correlated, simple regression can be used. When a clear relation exists between two independent variables and the residuals are small, the variables correlate and should not both be included in the model.

Linear/non-linear

The relation between dependent and independent variables can be described in different ways. When the dependent variable depends linearly on the model parameters (α_1 & α_2 in 3.1) the model is called a linear model. When the observational data are modelled by a function which is a non-linear (e.g. the dependent variable is related to the multiplication of two variables) combination of the model parameters the model is called non-linear. In between there is a log-linear model in which the variables are entered in the model as logarithms. This makes that normally non-linear models become (log-)linear models (like eq. 3.2) which makes them easier to work with.

Functional form

Also the functional form of the different regression models can differ. Both the normal or logarithmic values of the dependent and the independent variables can be used. Bresson et al. (2003) use a log-log functional form which means that the logarithm of both the dependent as the independent variable is entered in the model (like equation 3.2). The assumption that lies behind this is that the elasticity is constant over all levels of variables.

$$\log D_i = \alpha_1 \log I_{1,i}^{\beta_1} + \alpha_2 \log I_{2,i}^{\beta_2} + \varepsilon_i \qquad (3.2)$$

Kuby et al. (2004) however use a non-log function (comparable to equation 3.1) to describe the relation between the dependent and independent variables. Another possible functional form is the semi-log function, where some variables are not taken as logarithms (like equation 3.3).

$$\log D_i = \alpha_1 \log I_{1,i}^{\beta_1} + \alpha_2 I_{2,i}^{\beta_2} + \varepsilon_i \qquad (3.3)$$

Bresson et al. (2003) investigated whether using a semi-log model gave better outcomes than a log-log function and concluded that it gave better results in some cases and in other cases the result was equally good. If fare price, or a comparable variable, is taken into account it might hence be a good idea to check whether a semi-log model gives better results than a log-log model.

Estimation methods

The best model for the relation between the dependent and the independent variables has the smallest error terms (ε_i in equations 3.1-3.3). There are different parameter estimation methods that all work in a different way. The following are used by studies that are previously undertaken:
An estimator that is used frequently is the Ordinary Least Squares (OLS) estimator. The basic idea behind this estimator is that the sum of the squared differences between the model value and the observed value should be minimised. Kuby et al. (2004) use this estimator to give a first impression of the relations, reduce the number of independent variables and help identify heteroscedasticity. The absence of heteroscedasticity is a constraint when OLS is used. Bresson et al. (2003) also use the OLS estimator to determine whether coefficients were equal for all PT-areas in England and France.
Next to OLS the Instrumental Variable (IV) estimator is used by Bresson et al. (2003). This estimator is usually used when indirect relations are to be studied. Especially when the mediating variable cannot be measured itself. Just as with OLS the sum of the squared differences between the model value and the observed value should be minimised.
To correct the model for heteroscedasticity Kuby et al. (2004) use the LIMDEP statistical program. Their estimated model coefficients then did not change, but the standard errors did change (became smaller). LIMDEP is a program of Econometric Software, inc. and costs about $600[2].
Bresson et al. (2003) use shrinkage estimators to improve the estimated parameters. Shrinkage estimators are used to combine a naïve or raw estimate with other information. Because (root) mean square errors ((R)MSE) should be zero in perfect models the (R)MSE of the first-difference OLS or IV models should be combined with the "information" that it should be zero. Doing so should reduce the errors and improve the model. Bresson et al. (2003) use the mean absolute percent relative error (MAPRE) to give an indication of the goodness of fit of the model.

Assumptions

To make regression analysis work there are some assumptions that have to hold regarding the data that the model is estimated about. Two assumptions that are common to all estimation methods are:

- There have to be more research subjects (e.g. zones or stops) than variables. Otherwise it is not possible to estimate all the parameter values.
- The variable values of the different research subjects should be error free (without measurement errors), otherwise the estimation of the parameters is not right.

Other assumptions about regression models are the Gauss-Markov Assumptions (Kuby et al, 2004):

- Multi collinearity among independent variables, the independent variables are not correlated with each other. This can be checked by generating the Pearson correlation coefficient for all pairwise combinations
- Homoscedasticity, the variance of the residuals is independent of the value of the dependent variable, can be checked by plotting the errors against the value of the dependent variable
- The errors have expectation zero ($E(\varepsilon_i) = 0$)

Discrete choice models

Discrete choice models attempt to model the decision process of individuals or groups of people between alternatives.

Outline

The main idea behind discrete choice models is that the probability of individuals choosing a given option is a function of their socioeconomic characteristics, personal preferences and the relative attractiveness of the option (Ortúzar & Willumsen, 2001). To represent the attractiveness of an alternative the concept of utility (which people are thought to maximise) is used. In transport science discrete choice modelling is used for modal or route choice. The variables that determine the utility (U) of a car route can for instance be an alternative specific constant (ASC) the travel time (TT), the fuel costs (FC) and toll costs (TC), see below:

$$U_1 = ASC_1 + \alpha TT_1 + \beta FC_1 + \gamma TC_1 + \varepsilon_1 \quad ` \quad (3.4)$$

[2] http://www.limdep.com/products/limdep/

$$U_2 = ASC_2 + \alpha TT_2 + \beta FC_2 + \gamma TC_2 + \varepsilon_2 \quad (3.5)$$

The alternative specific constant is used to represent the net influence of all unobserved, not explicitly included characteristics of the individual. These can for instance be: peoples preferences for certain road types, peoples preferences for scenic beauty or peoples dislike of steep hills.
The modal choice step in the classical four-step model of travel behaviour is also an example of a discrete choice model. For all Origin-Destination pairs a set of alternative modes is available with certain characteristics, such as: travel time, costs and comfort.
The values of the parameters are usually estimated using maximum likelihood estimation.

Probability determining models
To determine the probability of people choosing a certain alternative, models are used. There are multiple models that are able to do so.

Multinomial Logit Model
Multinomial logit (MNL) models determine the probability of certain alternatives to be chosen based on the exponential utility in relation to the sum of the exponential probabilities of all alternatives (eq. 3.6).

$$P_i = \frac{\exp(\beta U_i)}{\sum \exp(\beta U_j)} \quad (3.6)$$

Special property of the model is the independence of irrelevant alternatives. This means that the relative probability of two alternatives being chosen is not different when the number of alternatives is changed (Haaijer, 1999).

Nested Logit Model
A nested logit model is comparable to a multinomial logit model but should be used when the independence of irrelevant alternatives property does not hold. This is the case when people only consider going by car or public transport. What kind of PT then is not interesting for them. When however bus and train are considered as separate alternatives in a MNL model, the total number of PT-trips will be too high. In a nested logit model therefore first the utility of a PT-trip is determined based on the bus and train trips (the lower nest). In the upper nest then that utility is used to determine the probability of people using car or public transport (Ortúzar & Willumsen, 2001).

Binary Probit Model
A binary probit model is comparable to a logit model, but the unobserved error terms are distributed standard normal rather than logistic. The function for the probability of an alternative (P_i) therefore becomes like equation 3.7.

$$P_1 = \Phi[(U_1 - U_2)/\sigma_\varepsilon] \quad (3.7)$$

Where:
Φ = the cumulative distribution function of the standard normal distribution
σ_ε = the spread of the errors
Binary probit models can be used when only two alternatives are considered (Ortúzar & Willumsen, 2001).

Elasticity
Elasticity is commonly used to describe the relations between independent and dependent variables. Elasticity tells something about how much people are changing their behaviour when a variable changes. The basic idea is that the elasticity is the relative change in the dependent variable divided by the relative change in the independent variable. If for instance the price of a bus ticket decreases with 10% and the number of trips increases with 3% the elasticity is -0.3 (+3/-10). Using elasticities to describe the relationships between the demand for bus transport and different variables is widely used in literature (see: Bresson et al, 2003, Holmgren, 2007 & Paulley, et al., 2006).
The common formula for the (point)elasticity, where the effect of variable x_i on the demand (Q) is described in terms of elasticities (E), is:

$$E = \frac{\delta Q}{\delta x_i}\frac{x_i}{Q} \quad (3.8)$$

The elasticity hence is dependent on the derivative of the function of the demand for bus transport over the variable and the value of both the variable and the corresponding demand. The elasticity therefore differs depending on where the point in question is located (Holmgren, 2007).

4 Research methodology

In the research methodology it is described how the actual research is performed. Subjects that are described are: how is the dependence measured, what spatial resolution is used, what to include in the research (project scope), what model to use and what time horizon to use.

4.1 Modelling dependence

The dependence of a dependent variable on different independent variables is represented by an equation that describes the dependent variable as a function of the independent variable(s) (D(I)).

4.1.1 Dependent variable definition

First it is important to define what the dependent variable is. Since we are interested in the demand for public transport this is the dependent variable. The "demand for public transport" itself is however quite abstract and therefore it is represented by another variable.

The demand for public transport can be measured in different ways:

- The number of journeys per capita (Bresson et al, 2003, Fitzroy & Smith, 1998 & Souche, 2010)
- The number of boardings per zone/stop (Kuby et al, 2004)
- The number of passenger kilometres

Most known literature uses the number of journeys per capita as the measure for the demand for public transport. For a transport region as a whole this is easy to obtain because you just need to know how many people live in the area and how many trips are made in a certain period in the region. If you want to compare the different zones that are in a certain region you have to do a survey that asks people how many times they travel by bus. Because you need a representative sample for every zone, the acquiring of these data takes quite some effort. The alternative is using data of the Dutch Mobility Research (MON). On a household/personal level, which is needed to compute the number of trips per capita, only data on a municipal level is known in that dataset. Using these data therefore would take quite some effort.
Another possible measure of the demand for public transport can be the number of boardings per zone or stop. In this way all trips that are taken can easily be assigned to a certain zone. There are however also some problems with this measure. Because many people might use the bus as a complement to the train many people enter the bus in the zone where a train station is. In that way that zone/stop has a high "bus transport demand" while only some of the boarders live in that zone. This can be overcome when you only look at home based trips. Another possibility is to introduce a dummy variable (which has a value of either 1 or 0) for the presence of a train station, like Kuby et al. (2004) did for the effect of an airport on light-rail passengers. You can also assign trips from the rail station to the zone where people exit the bus. Another possibility is not taking into account the zone with the train station. The MON has trip data on a Postal Code-4 (PC4) level[3]. With these data it is possible to give an impression of the number of bus trips from a certain zone, but the quantity of data is not enough to give reliable trip numbers. At Regio Twente NVS passenger counts on certain points of all lines are present (Connexxion, 2008). These can also be used to determine how many people have boarded in between the points. Because these counts were not performed at all stops it is impossible to make boarding per stop data from these counts.
These two measures don't consider the distance that is travelled by individual persons. Because this distance is important for the price people pay for the service it might be a good idea to look at the passenger kilometres travelled from each zone. Problem with this measure is that zones that lie further away from the centre get a higher level of demand for bus transport than zones close to the centre with equal numbers of travellers. Therefore this is not a useful measure for the demand for bus transport.
Because most data are available on a zone level (see chapter 7 about data) the research takes the number of boardings per zone as the representative for the demand for bus transport in that zone.

[3] The postal codes in the Netherlands consist of 4 numbers and 2 characters (i.e. 7500 AB). In a Postal-Code 4 area all households have the same number combination while they can have different characters. A PC-4 area hence is larger than a PC-6 area.

4.1.2 Timescale

Demand is not constant over a year. In holiday periods the demand for transport for instance is usually lower because fewer people have to travel to work and school. In a week the demand is also not constant while people usually travel less in weekends and least on Sundays. Also on a time-of-day level the demand changes from time to time (during peak hours the demand is higher than outside peak hours). The characteristics of the different time periods are different and it is therefore important to determine what timescale is used for the research (boardings per year, per week, per workday or per peak-hour).

One often used time scale is an average morning peak hour or period. In this way mainly trips from home to work and school are included in the study. The advantage is that the origins of most of the trips are the houses of people which makes the trips easily attributable to a certain (origin) zone. In this way also the previously mentioned problem with high demand in areas with a train station might also become smaller. Disadvantage is that trips with other purposes, where elasticities are higher (Paulley, et al., 2006), are not really taken into account in this way.

Using an average workday (Monday-Friday) is another interesting possibility because the NVS-counts that Regio Twente has are average passenger numbers for 10 workdays. Advantage of using this timescale is that leisure and social trips are better taken into account (during peak hours these trips are undertaken relatively less). In this way also the trips that are more elastic are taken into account.

The timescale for this research is an average workday because traveller data about the average workday are present and also trips back to home are taken into account.

4.2 Spatial resolution

The spatial resolution which is used for the research is dependent on two important factors, the detail that you want to have the outcomes (as high as possible) and the effort that has to be made to obtain and edit the data (as little as possible). There has to be a balance between these two variables.

4.2.1 Level of detail

Ideally a model can simulate the mode choice process for all individual people exactly. To do that for all individual people all kinds of personal information has to be available about all people. For a city like Enschede alone there have to be more than 155.000 research objects. The effort it costs to make a data file with over 155.000 records is off course huge. Therefore this level of detail is not feasible.

Another extreme model would just have one zone and consider the averages of the whole study area on different variables, which is easy to obtain. It is obvious that this is also not a feasible model because differences between the neighbourhoods in the study area, which are needed to answer the research questions, then cannot be detected. The model hence has to be in-between the two mentioned models.

There are some constraints, which follow from the research questions, that have to be kept in mind. They are:

- The model should be made with data of different zones/stops (as many as possible)
- The zones/stops all need to have a certain (average) value for the different possible variables
- The spatial scale of the different data sources should be able to be matched with each other

4.2.2 Data availability

There are different organisations that collect and manage data from the study area. Statistics Netherlands (CBS) for instance collects data from all over the country. Other organisations have (more detailed) data from smaller regions. A more detailed overview can be found in chapter 7. In this chapter it is concluded that most data are available at a neighbourhood level. The number of boardings however is only known at a zone level (one zone is a group of neighbourhoods).

4.2.3 Conclusion

It can be concluded that there are many data available on a PC4 level[4]. Therefore this spatial scale seems to be a good one. There are however also many data on a more disaggregated level (some even at a bus-stop level). Aggregating these data to PC-4 areas could mean a substantial loss of inter-zonal differences. Therefore the zone level is used, in which one zone is built up out of multiple neighbourhoods.

Some data are however available at larger or smaller scales. Data about larger zones are (dis)aggregated in on of the following ways:

[4] A Dutch postal code (PC) contains 4 numbers and 2 letters (like 7511 JM). PC4 areas are zones that all have the same 4 numbers in their postal codes.

- Assuming the variable to be constant over the smaller zones. If job density for instance is only known on a PC4 scale, it can be assumed that the job density in all (smaller) zones of that area the job density is equal to that of the whole area.
- Divide the value based on another variable. When for instance the number of shops is known on a PC-4 level while the number of jobs is known on a smaller scale the number of shops can be assigned to the smaller scale zones using the same relative distribution as the jobs. If 25% of the jobs in the zone are in a certain sub-zone the also 25% of shops are allocated to the sub-zone.
- Averaging the values in the smaller neighbourhoods with respect to the number of inhabitants of the areas of the neighbourhood.

Disaggregating the data is clearly described and underpinned to make sure the outcomes of the research remain valuable.

4.3 Project scope

Because the time for this research is limited and data collection can be very time consuming it is important to determine what to study and, more important, what not to study. It is impossible to study all possible determinants in the whole concession area of Twente. Therefore the research is demarcated here. The research area, modes and variables to study are all described separately.

4.3.1 Research area

As mentioned before, studying the whole concession area would be too extensive for this research. Therefore only three municipalities (Enschede, Haaksbergen and Losser) are taken into account. In this way an urban area (Enschede), a rural area with many bus lines and stops (Haaksbergen) and a rural area with less lines and stops are taken into account. When the relations for the different municipalities are known other municipalities can be categorised as one of the other three to tell something about the demand for bus transport in that municipality.

4.3.2 Modes

Another important demarcation of the project scope is that only bus transport is taken into account and not other kinds of public transport (like train). A reason to exclude the train is that it is used for other kinds of trips than the bus. Bus trips are for instance (on average) much shorter than train trips (Rijkswaterstaat, 2008). In many cases a (regional) bus network serves as a feeder for the (national/inter-regional) train network, which means that many train trips have an origin in another (concession) region than they have their origin. This makes the train travel more difficult to trace and also less interesting for Regio Twente. Including train trips would also increase computational complexity hugely. Therefore train trips are not taken into account. It is however important to note that many bus trips go to (areas around) train stations.

4.3.3 Selection of variables

It is impossible to include all in paragraph 3.3 mentioned variables. Therefore a selection is made of the variables that are said to be most important and that are easy to obtain.

Socio economic

Probably the most mentioned socio economic variable that influences the demand for bus transport is car ownership. The idea that the more cars there are the less the bus is used for transportation seems very logical. Even more logical seems that when people do not have a car they travel by bus more often. Hence car ownership is an important determinant of the demand for bus transport. The average number of cars per inhabitant in a zone is the measure of car ownership.

Also business car ownership is included, because it was proven that people that own company cars use cars more often. It seems logical that these people use busses less. The measure for this variable is the number of business cars per inhabitant.

Although income does influence car ownership, it is included in the research. The reason for that is that it is interesting to study whether income has only an indirect effect on bus use, like with car use (Van Acker & Witlox, 2010), or also a direct effect. If this is the case, people with different income levels that both do (not) have a car have different bus ridership numbers. The measure that is used for the income is the standardised income per household. Because employment status and educational level are highly related to income these two variables are not included in the research.

Because students are thought to use busses more often students are also included in the research. The expectation is that the more students there are in a neighbourhood the more bus demand exists in the neighbourhood. Ideally the number of students in each zone is used as a measure, but because

these data can be more difficult to obtain also the number of people aged 15-24 is used as a measure for the number of students.
The other variables that were mentioned earlier (like: age, gender, culture, renter) are mainly determinants for the other variables and are therefore not included in the research.

Built environment
The model that is currently in use for transport modelling uses the number of jobs and inhabitants to determine the number of trips from and to a certain zone. The number of bus boardings is also influenced by the number of inhabitants in a zone. The basic idea is that a neighbourhood with 100 inhabitants/jobs has less boardings than a neighbourhoods with 1500 inhabitants/jobs. These variables are known, which makes using another measure not needed.
To include the accessibility of destinations by bus five possible measures are included in the research. These are the relative job accessibility by car and bicycle, the travel time by bus towards the city centre and the relative travel time towards the city centre by car and bicycle.
Not all trips are home-work trips. It therefore might be needed to include another measure for the relative attractiveness of the different modes. Because people also transport to visit other people the number of inhabitants that are accessible by a mode within a certain amount of time can also be important. Because the number of jobs in a zone and the number of inhabitants in a zone are highly correlated (see paragraph 7.5.3 and appendix C) only the relative number of accessible jobs is included in the research.
Another important determinant is the spatial density in an origin zone. The higher the density the more people live/work near a bus stop and the more they are thought to use the bus. The measures of this variable are population density, job density and address density. Job density however is also related to the number of jobs.
Where students live determines to a large extend where students enter the bus system on their way to college. Where they enter the bus system on their way home is mostly determined by the location of their college. Therefore the number of college enrolments is a built-environment related variable that can be used to include education-home trips. The education organisations that are included are, the University of Twente, Saxion Hogescholen, ArtEZ and ROC Twente.

Network
The frequency of services at a stop is a very important determinant for the people that travel by bus in a zone or that enter the bus on a certain stop. Therefore the frequency is included in the research. Because a whole workday is the timescale the measure for the frequency is the total number of services at a stop on a working day.
The travel time towards a stop (egress time) also is included in the research. Because a zonal scale is used it is interesting what the average distance to a stop is in that zone.
Because punctuality data are available at Regio Twente the punctuality of the different lines is also taken into account for the research. Theoretically lines with a higher punctuality should attract more passengers than lines with lower punctuality. The percentage of services that departs less than 1 minute early and leave less than 3 minutes late is the measure for punctuality.
A special variable that is related to the number of boardings is the presence of an important train station. When there is a train station the number of bus boardings is substantially higher than when there is no train station. In the study area there are four train stations of which one (Enschede Central station) is the station where (nearly) all lines stop. The stop that is near that intercity station therefore has a relatively very large number of boarding. To account for the extra travellers a dummy variable is introduced. This variable is only true for the central bus station in Enschede.

Personal preferences
Currently no research has been done on the personal preferences of people living in the study area. Including these variables in the research would therefore cost a lot of time because large amounts of surveys, that are representative for all individual zones, should be undertaken and analysed. Therefore it is decided to keep these variables out of the project scope of this research.

Overview
An overview of the variables that are included in the research and the relations between them is given in figure 4.1.

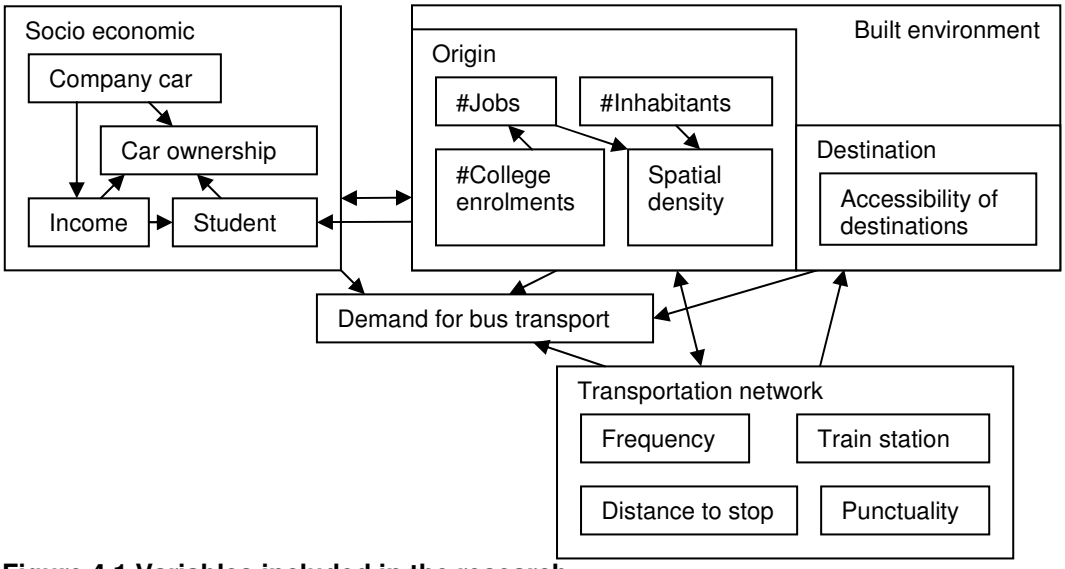


Figure 4.1 Variables included in the research

4.4 Model

Since the variables that are studied are known it is important to determine how the study is performed. The type of analysis method and the model characteristics are the two topics that are of interest. First however the currently used car traffic model is described briefly.

4.4.1 Current model

The model that is currently in use by Regio Twente is the Regional Traffic Model, that makes use of Omnitrans. This is a model that is developed by the Goudappel Coffeng, a Dutch transport and traffic consulting firm. The base of the model is the road network of the Netherlands. With help of three of the four steps of transport modelling (Ortúzar & Willumsen, 2001) the model determines the number of vehicles on certain links. Public transport is not included in the model. On that road network zones are connected that all have a number of inhabitants and a number of jobs. The model generates an Origin-Destination matrix that describes all trips in a peak period in the area based on the inhabitants, jobs and distance between the zones. Then these trips are assigned to the network based on the shortest travel time between the origin and destination zones. With help of traffic counts the model is then calibrated to come to a better representation of the true traffic numbers (Friso & Van der Honing, 2006).
The model is used to give insights in future traffic numbers and to calculate what the effects of changes in the road network are. In this way it can be tested what effects new infrastructure has without having to construct the link first. Also insight can be obtained on areas that can become highly congested if nothing is changed on the road network (Friso & Van der Honing, 2006).

4.4.2 Type

In paragraph 3.2.2 two different types of analysis models were described: regression models and discrete choice models. Discrete choice models are used to determine modal choice for individual OD-pairs. Because this research looks at certain zonal or stop-level characteristics and only takes accessibility to jobs as a variable that is related to the destination of a trip a discrete choice model is not a useful analysis method for this research. Regression analysis on the other hand is used to describe relations between certain characteristics. Therefore this method is chosen for this research.

4.4.3 Functional characteristics

The first functional characteristic that has to be decided about is whether all variables are modelled independently (simple regression) or simultaneously (multiple regression). Because it would be ideal to come up with a model that can determine how many boardings one can expect given certain (multiple) characteristics multiple regression is chosen. Correlation between independent variables are checked to make that highly correlating variables are not included in the model together.

The second functional characteristic is its form. Because the different variables should not be related to each other the model that is estimated is a linear model. Within this category it is possible to have a normal form (D(I))[5] a log-log form (logD(logI)) and semi log forms (logD(I)/D(logI)). A study by Kuby et al. (2004), that is comparable to the research that is performed, used a normal functional form. Therefore this research also uses a normal linear functional form.
The last functional characteristic that is defined is the estimation method that is used to estimate the parameters of the model. Methods that are used in literature are: OLS and IV. IV is usually used when the relation between a mediating variable that cannot be measured and a dependent variable has to be determined. The variables that are included in the research are however measured themselves, which means that IV is not the designated method. Because OLS is a simple method that can be used to give a first impression of the relations, this method is used for the research.

4.5 Time horizon

It is important to think of how many years are studied. If you study only one year it is possible to give a very detailed overview of the differences in demand between the different zones. When however multiple years are studied differences between the years (changes in price of tickets or frequency) can be taken into account for the same zones. Disadvantage of the time-series analysis is however that there is more data needed. It therefore depends on the availability of the data whether a time-series analysis can be performed within the time constraints of the research. In the research first a model is fitted to the data of one year (2008). When that is done the model is validated using another year (2004). Of those years one lies in the period of the current concession period and the other in the previous one to check whether the model can predict the differences in the number of boardings because of the changes (mainly in frequency) that have been implemented at the start of the new concession.

[5] Where: D = the dependent variable, I = the independent variable

5 Research Area

In this chapter the research area is described in detail. Also the different lines in the research area are introduced.

5.1 Location in Twente

The three municipalities that are part of the research area are located in the South-Eastern part of Twente (see figure 5.1).

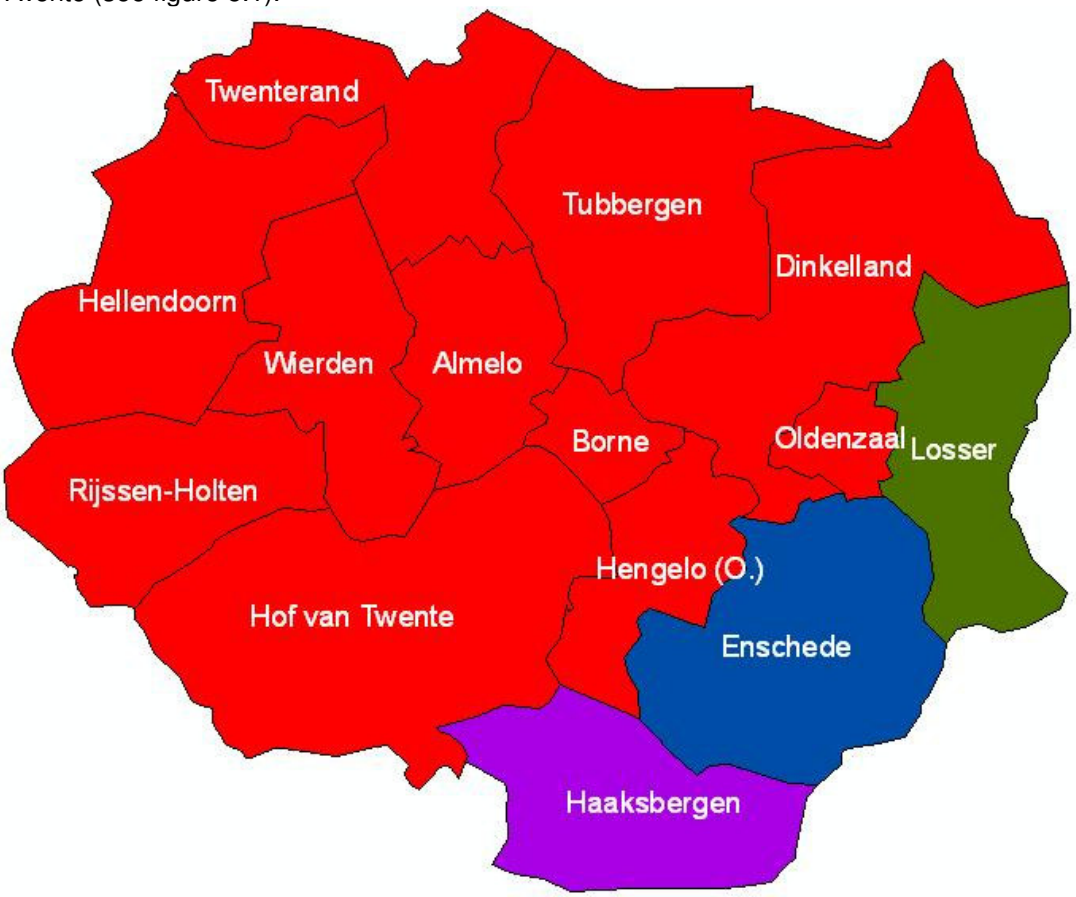


Figure 5.1 The research area in Twente

5.2 Enschede

Enschede (figure 5.2) is the largest city in Twente and has about 157.000 inhabitants. In the city many national/regional facilities are located such as a university, institutes for professional education, one intercity station and three suburban train stations, a hospital, cultural institutions, sports venues and a wide variety of shops and services. Because of these facilities many people travel to Enschede for work and leisure purposes. Because Enschede is a regional centre also many business areas are located around Enschede (mainly around the harbour in the west of the city).
In the municipality also some smaller villages are located such as Glanerbrug (east of Enschede along the national border), Lonneker (north) and Boekelo (west).
Enschede has an extensive urban bus network that services the whole city and both Glanerbrug and Boekelo. The lines of this urban bus network are Line 1, 2, 3, 4, 5, 6 and 7. Besides these lines also regional lines are exploited to Hengelo (9, 15, 16, 17 & 18), Oldenzaal (60 & 62), Overdinkel (61), and Haaksbergen and beyond (20, 73, 74 & 76).

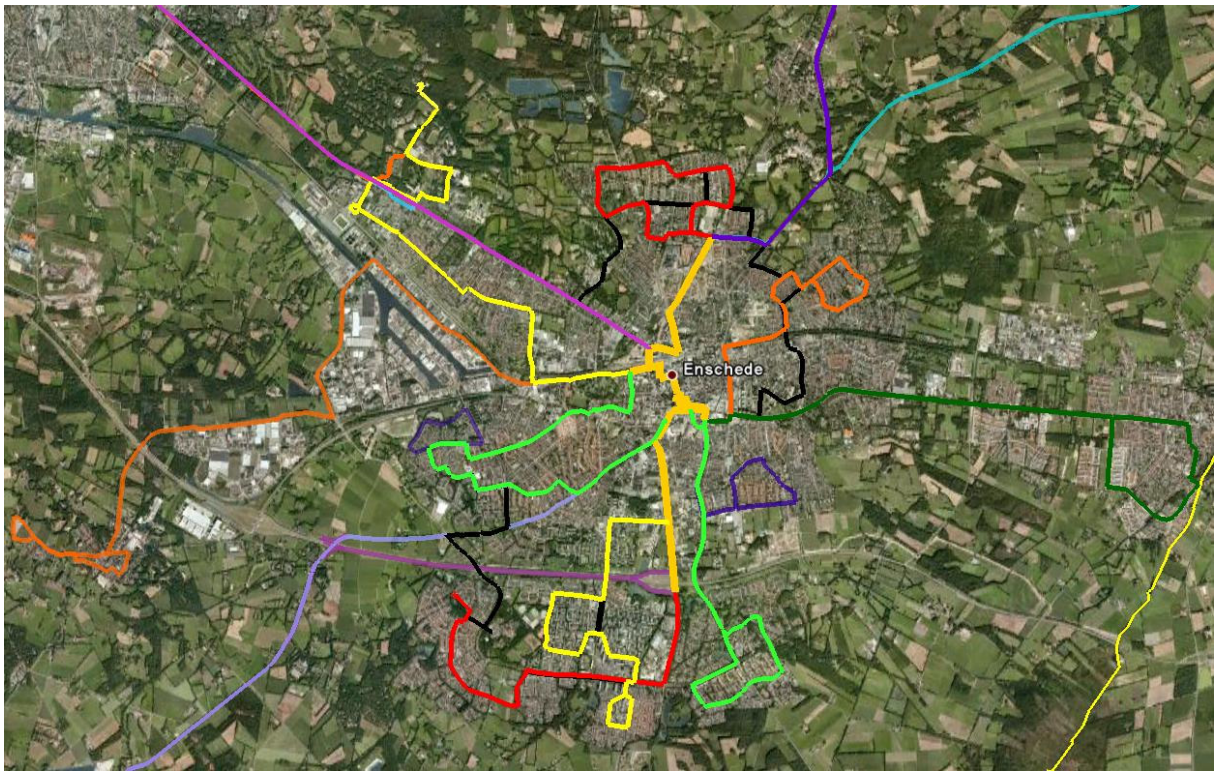


Figure 5.2 Map of Enschede

5.3 Haaksbergen

Haaksbergen (figure 5.3) is a rural municipality including one relatively large village (Haaksbergen) and two smaller villages (Buurse and Sint-Isodorushoeve) with nearly 24.500 inhabitants. Haaksbergen has one train station that is only in use for leisure trips, an affiliated rail museum and some small business areas in the West and North.
Through the municipality many lines run. Four lines (20, 73, 74 and 76, Purple) go to Enschede, two (53, Green, and 59, Blue) to Hengelo and one (97, Red) to Goor. In Haaksbergen these lines run criss-cross through the village which makes the spread of the stops very good.

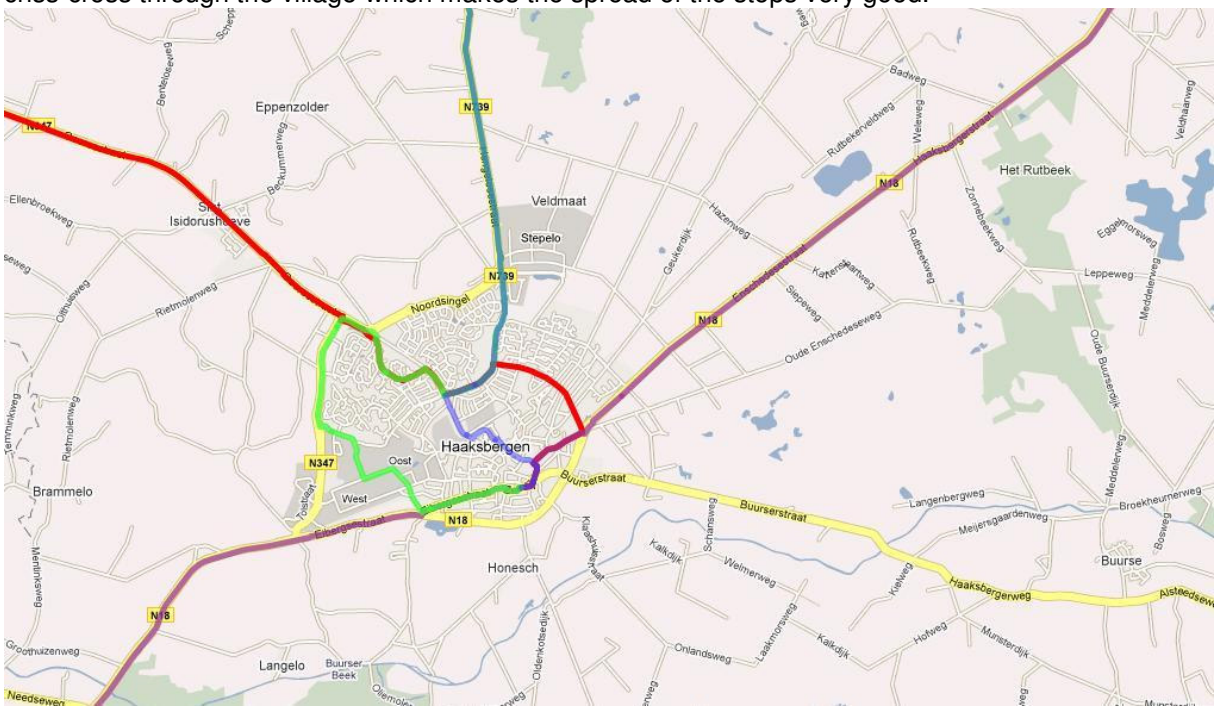


Figure 5.3 Map of Haaksbergen

5.4 Losser

Losser (figure 5.4) is another rural municipality with about 22.500 inhabitants. Within the municipality also the smaller villages Overdinkel, de Lutte, Glane and Beuningen are located. Through the municipalitiy three normal buslines are operated (61, 62 and 64). Only two of them really service the villages in the municipality: Line 61 (Blue) Overdinkel-Enschede vv and line 64 (Orange) Overdinkel-Oldenzaal vv. The lines run shrough the centres of Overdinkel and Losser and than split to Enschede and Oldenzaal. Because the stops are only located in one line through the centre the spread of the stops is worse than in Haaksbergen.

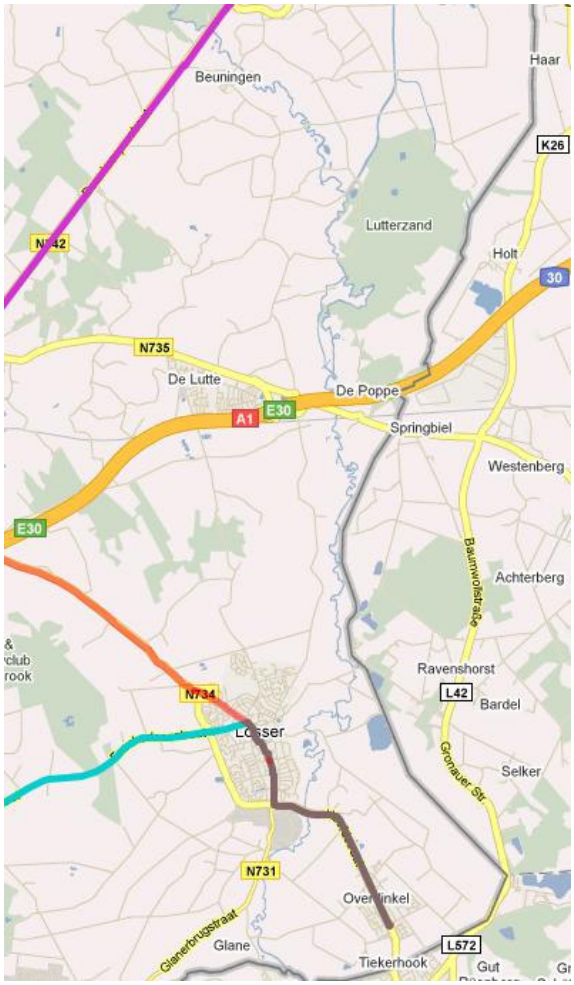


Figure 5.4 Map of Losser

6 Zoning

Because of the scale level of the number of boardings a subdivision of the research area into zones is needed for the research. The size of these zones depends on the lines that run through the neighbourhoods of the different municipalities. In this chapter the subdivision into zones is described with a special focus on the allocation dependency to the bus lines.

One of the subdivisions that are already present is the subdivision in neighbourhoods that is used by Dutch Statistics. These neighbourhoods can however not be used because the spatial resolution of the boarding data is larger. Therefore the research objects are formed by groups of neighbourhoods. An overview of the different zones and the corresponding lines and neighbourhoods is given in appendix B. In figure 6.1 the different colours indicate the different zones.

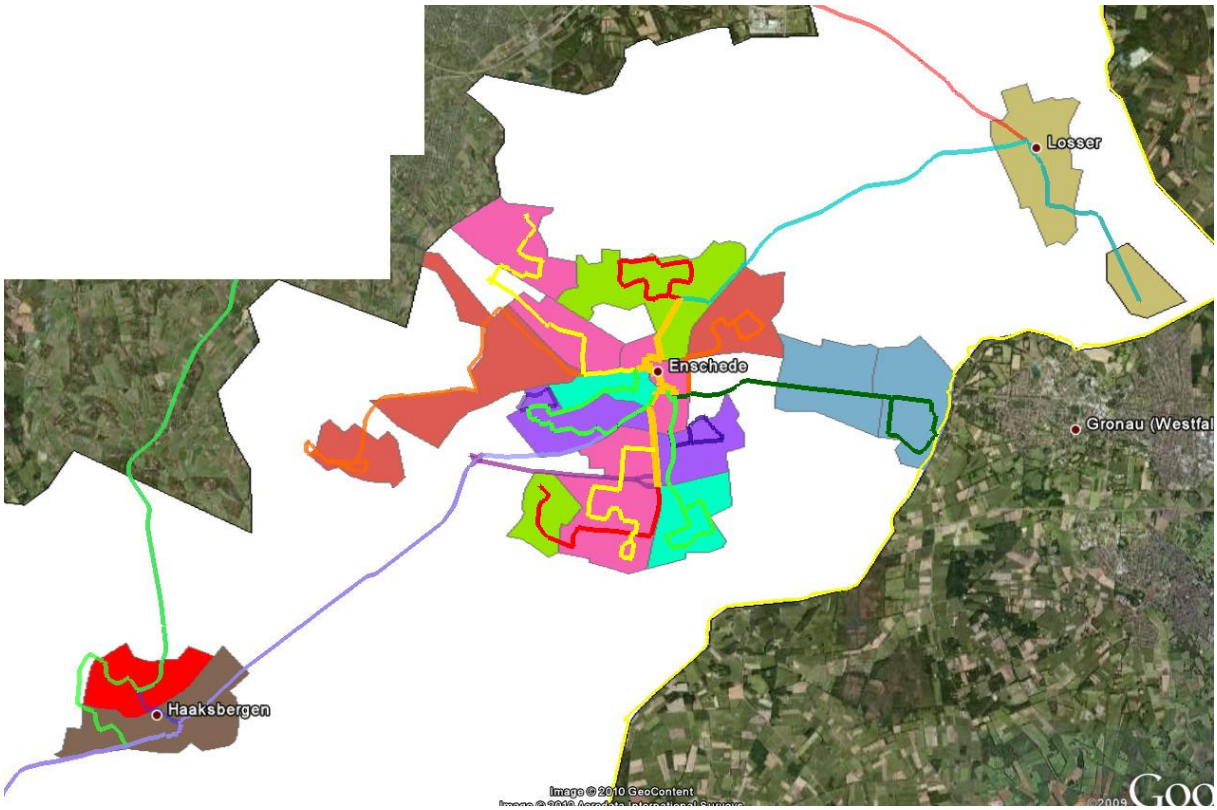


Figure 6.1 Groups of neighbourhoods as research objects

6.1 Subdivision

The size of the zones is dependent on the line that the zone is related to. Line 1 for instance is build up out of 5 zones, from North to South: The university, Twekkelerveld, The city centre (nearly all lines have a zone here), Boswinkel and Wesselerbrink (see figure 6.2).

Figure 6.2 Zones that are serviced by line 1 (Pink)

A neighbourhood either is fully part of a zone (the research object) or not at all. Neighbourhoods are not subdivided further.
The basic assumption is that all people are travelling to the city centre (including the train station). Because of that only the lines directly towards the city centre of Enschede are subdivided into zones that are included in the research. The service lines (7a and 7b) are not included because they make a large loop through Enschede. Other lines that are not included are the lines between Enschede central station and (neighbourhoods of) Hengelo (9, 17 and 18) and lines from the university to the train station of Hengelo (15 and 16). In Haaksbergen and Losser some lines do not go to Enschede, but to Hengelo or Oldenzaal. Because different destinations are serviced by the lines also multiple zones are formed from them.

6.2 Neighbourhoods serviced by multiple lines

Some neighbourhoods are serviced by multiple bus lines with either the same or different destinations. These neighbourhoods are described below.

Enschede
In Enschede some neighbourhoods are serviced by multiple lines. Some of the neighbourhoods of the Wesselerbrink district (South) are for instance serviced by both line 1 and 2 (figure 6.3). Two neighbourhoods therefore are part of two zones, the pink zone in figure 6.3a and the green zone in figure 6.3b).

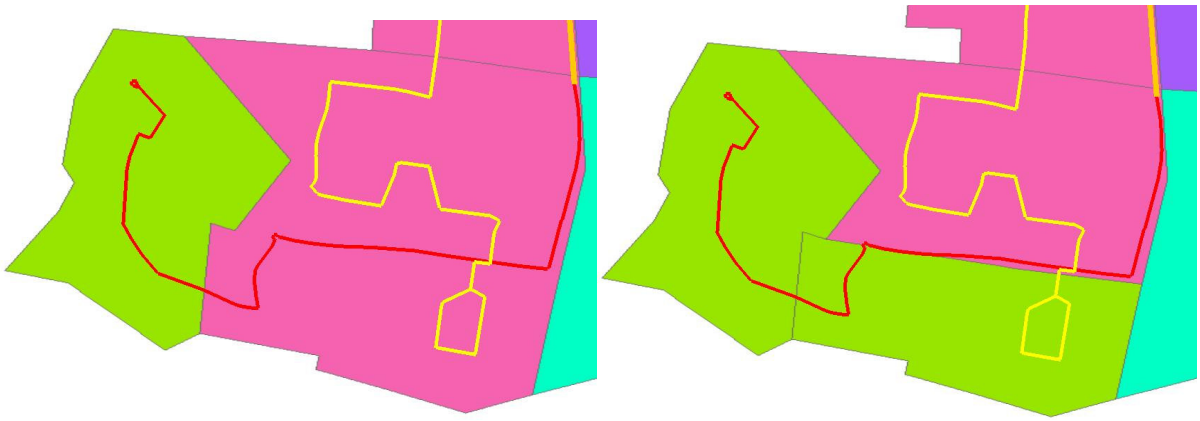


Figure 6.3 Line 1 (yellow) and 2 (red) through the Wesselerbrink

In the West of Enschede two lines are serviced in opposite direction. Together the lines make a loop through the western neighbourhoods (figure 6.4). Line 4 in the clockwise direction and line 5 in the counter clockwise direction. Because the assumption is that all people travel towards the city centre the neighbourhoods are assigned to the line that has shortest service towards the centre of Enschede. The green areas in figure 6.4 are thought to use line 4 and the blue neighbourhoods are thought to use line 5. One neighbourhood, farthest from the centre, is serviced by both lines with a reasonable equal travel time to the city centre. Therefore this neighbourhood is part of both zones (see the difference between figure 6.4a and 6.4b). The variable values of the zone therefore are also used in both zones.

Figure 6.4 Line 4 (green, clockwise) and 5 (blue, counter clockwise) through the western part of Enschede

Haaksbergen

The different lines through Haaksbergen are either towards Enschede (20, 73, 74 and 76) or Hengelo (53 and 59). Some of the neighbourhoods in Haaksbergen are only serviced by the lines towards Enschede, others only by the lines towards Hengelo and yet others are serviced by both groups of lines. In figure 6.5 the neighbourhoods that have access to Enschede are grey and towards Hengelo red. Altogether Haaksbergen is split into two research zones with some neighbourhoods being part of both zones.

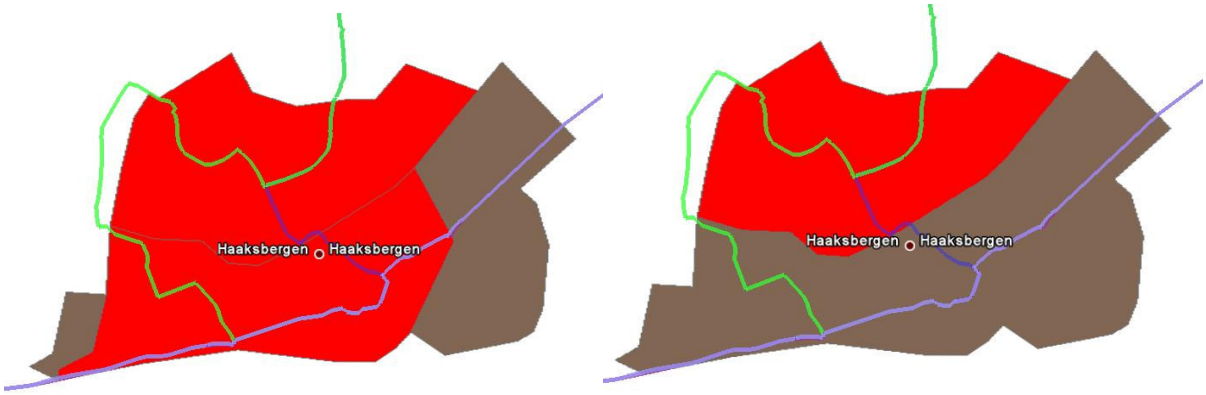


Figure 6.5 The centre of Haaksbergen is part of the zone with services to Hengelo (red) and Enschede (grey)

Losser

In the municipality of Losser also multiple lines are present with different destinations. Line 61 towards Enschede and line 64 towards Oldenzaal are both attributed to a different zone. All neighbourhoods of Overdinkel and Losser however are serviced by both lines which makes both zones equal. The number of boardings and the variables of line specific variables are however different.

7 Data

In this chapter the data that are used are described. After an overview of the different available datasets the different sources of the data are described. After that the different possible measures for the dependent variable are described and one of them is chosen. After that the different independent variables (socio economic, built environment and transportation network) are described. An important data source, the Neighbourhood map, is described in the last paragraph.

7.1 Overview

In table 7.1 an overview of the datasets that are available is given. In the table for every variable the source, the level of detail and the years for which the datasets are available is given.

Variable	Source	Level of detail	Years available
Boardings	PT-chipcard SLO NVS	Stop, all lines Stop, line 2, 20, 53, 59, 64, 73 and 74 Groups of neighbourhoods	From Jan-2010, monthly 2007 2004-2009
(Business) car ownership	Dutch Statistics	Neighbourhood	2003-2009
Income	Buurtmonitor Dutch Statistics	Neighbourhood (only Enschede) Neighbourhood (only household)	2007 2004
Number of students People aged 15-24	DUO Dutch Statistics	PC-4 Neighbourhood	2010 2003-2009
Inhabitants Population density	RVM (Regio)	Sub-neighbourhood	2004-2008
Jobs	RVM	Sub-neighbourhood	2004-2008
College Enrolments	University Saxion ROC ArtEZ	Point	1964-2008 2007-2008 2001-2009 2002-2008
Relative accessibility	RVM	Sub-neighbourhood	2008
Travel time towards centre	Connexxion Google maps	Stop Neighbourhood	2004-2010 2010
Job/inhabitant density Level of urbanisation	RVM Dutch Statistics	Sub-neighbourhood Neighbourhood	2004-2008 2003-2009
Address density	Dutch Statistics	Neighbourhood	2003-2009
Frequency	Connexxion	Stop	2004-2010
Punctuality	SabiMIS	Line	2006-2010, quarterly
Train Station		Point	
Distance to stop	Connexxion ArcGIS	Point (location of stops) Neighbourhood	2008

Table 7.1 Overview of the available datasets

7.2 Sources

The sources of the datasets are described here.

7.2.1 Concession holder

The concession holder for Twente (Connexxion) has information about different variables that are included in the research. They for instance have data about the location of the stops they service and about the number of services per stop (frequency).
Besides this information there are also datasets about the number of travellers on their busses. One of these datasets contains the number of passengers that are in the bus on some specified stops (about 3 or 4) on a line that are counted by the bus drivers. These counts are done in two weeks in November in such a way that the number of passengers on an average workday (with a total of 10 measurements) and on a Saturday and Sunday (n=2) can be estimated. In the dataset for every

scheduled trip the number of passengers inside the bus at the specified stops is shown. The trips are subdivided into weekdays, Saturdays and Sundays. Also the total number of passengers on an average (week-/Satur-/Sun-)day is included. An example is shown in figure 7.1.

Tabel	Bezetting per telpunt / rit					
Bron	NVS 2008 Najaar connexxion					
Lijn	Universiteit - Wesselerbrink					
Sys-lijn	D001					
Richting	1					
Dagsoort	MAVR					
Gemiddeld						
Nr.	Telpunt	**149 21:54 N= 10 45 zitp.**	**151 22:24 N= 10 45 zitp.**	**153 22:54 N= 10 45 zitp.**	**155 23:24 N= 10 45 zitp.**	
7	Enschede, UT/Bosweg	5,8	2,0	2,7	4,3	753,9
15	Enschede, Lambertus Buddestraat	11,9	7,7	6,0	6,4	1222,9
21	Enschede, MST Haaksbergerstraat	15,4	14,6	8,9	6,6	1299,3
27	Enschede, Tuindorp Broekheurne	10,8	6,3	6,7	3,0	772,8

Figure 7.1 Number of passengers on line 1, direction 1, average weekday on four counting points

The concession holder also monitors the number of passengers on specific lines using Line Researches (Dutch: Standaard Lijnenonderzoek, SLO). These are surveys that are undertaken by consultancy firms that question most of the passengers on a line at a certain day. The questions that are asked are about origin and destination, personal characteristics, travel characteristics and perceived quality.
Since the PT-chipcard is put into service also data about the (de)boardings of passengers are known for all lines and stops. The concession holder processes these data into a report that is send to the concession granting authority. The information is about: The number of (de)boardings per stop, the number of (de)boardings per line per hour, the number of passenger kilometres per line and for some stops the punctuality.

7.2.2 Regio Twente

Regio Twente also owns data about the research area. The Regional Traffic Model Twente (RVM Twente) for instance is a model that is used to calculate the effects of changes in demography and infrastructure on the travel behaviour of people in Twente. For this model the region is subdivided into many small zones that all have a certain number of inhabitants and jobs. Also the bus stops are included in the model. The model is developed by Goudappel Coffeng in close cooperation of the fourteen municipalities of Twente and Regio Twente. In figure 7.2 the spatial scale level of the RVM is shown.

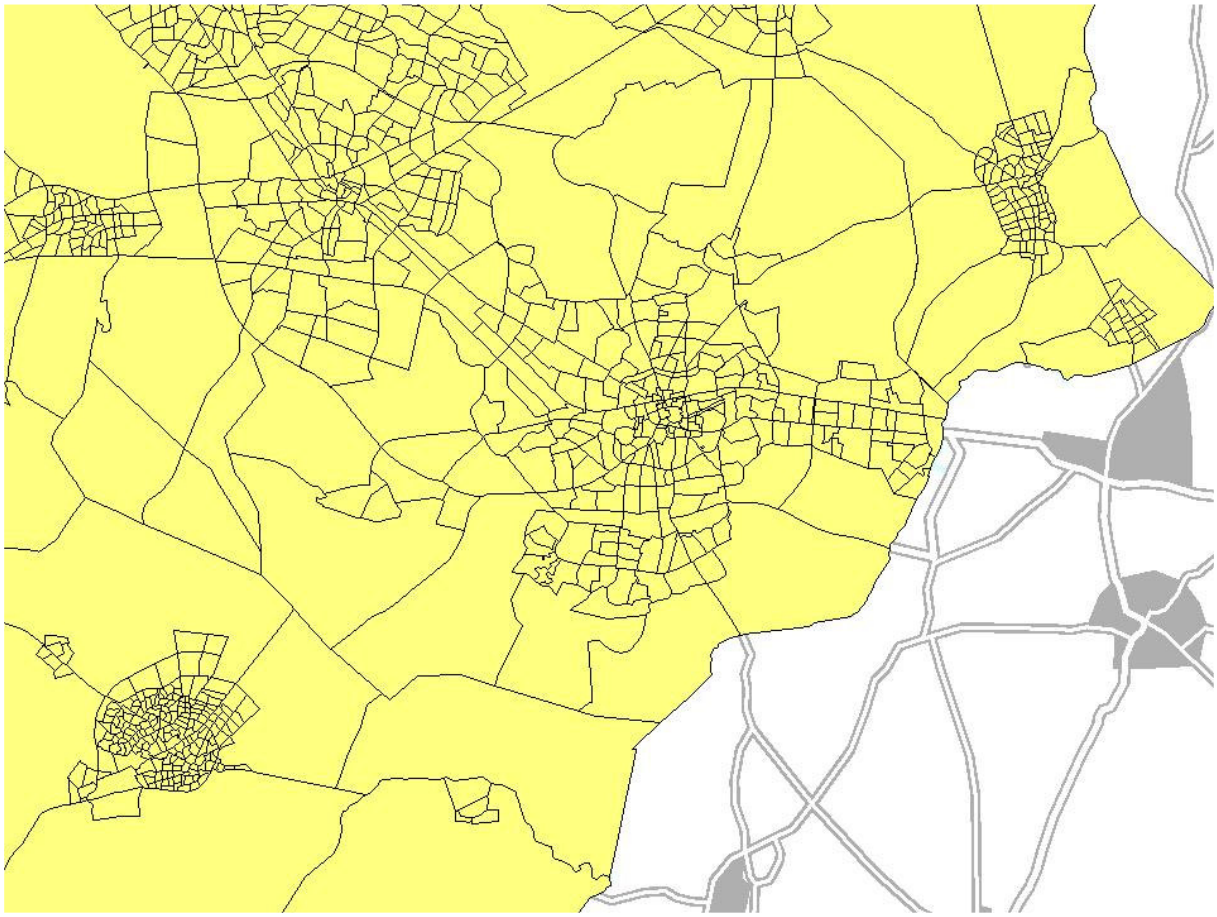


Figure 7.2 Spatial scale Regional Traffic Model RVM

7.2.3 Municipality Enschede

The municipality of Enschede has, unlike the other municipalities, detailed information available about different neighbourhoods in the city. This dataset is called the Neighbourhood Monitor and can be accessed online[6]. All kinds of information about the inhabitants, the economy, income, liveability, labour, dwellings and others are available at this website.

7.2.4 Dutch Statistics

One important source for the research is the Neighbourhood map that is produced yearly by Dutch Statistics. This map consists of the geometry of the Netherlands in three different scale levels: a municipal, district and neighbourhood level (figure 7.3). On these levels all kinds of information about inhabitants, age groups, minorities, surface area, income, car ownership, real-estate value, social security and businesses.

[6] http://enschede.buurtmonitor.nl/

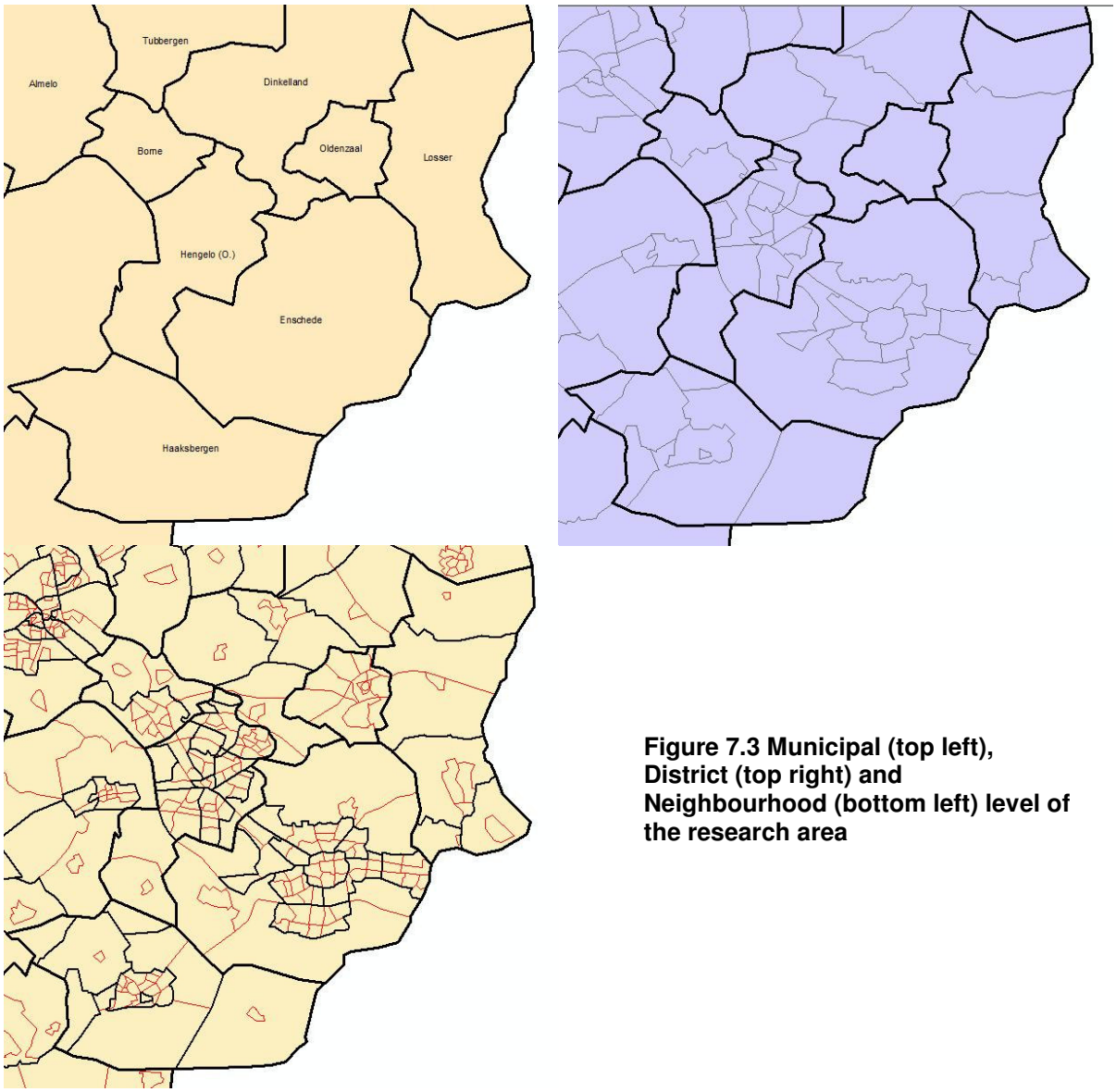


Figure 7.3 Municipal (top left), District (top right) and Neighbourhood (bottom left) level of the research area

Next to the Neighbourhood map that is published yearly also incidental numbers are published by Dutch Statistics. The standardised income per household in all neighbourhoods in the Netherlands in 2004 for instance can be found on the website of DS. The data come from a survey (regional income distribution) that is performed by Dutch Statistics. The error margins are reasonably high (1,500 euro for a zone of 200 inhabitants and 700 euro for a zone of 300 inhabitants). Per household the absolute error margins are even larger (3,100 and 1,900 euro). Reliability hence is not perfect, but better data are not available.

7.2.5 Educational institutions

The educational institutions have information about the number of students or scholars that they educate and the location of their buildings.

7.2.6 OCW DUO

OCW DUO is the new name for the IB Groep. This is the institute that registers all students and also coordinates the payment of the student financial aid and the delivery of the Student-PT-(chip)card. Because the addresses of all students are known it is possible to determine the number of students at a PC-4 level. More detailed information is, because of privacy issues, not delivered by OCW DUO. As an example the PC-4 areas in Enschede are shown in figure 7.4.

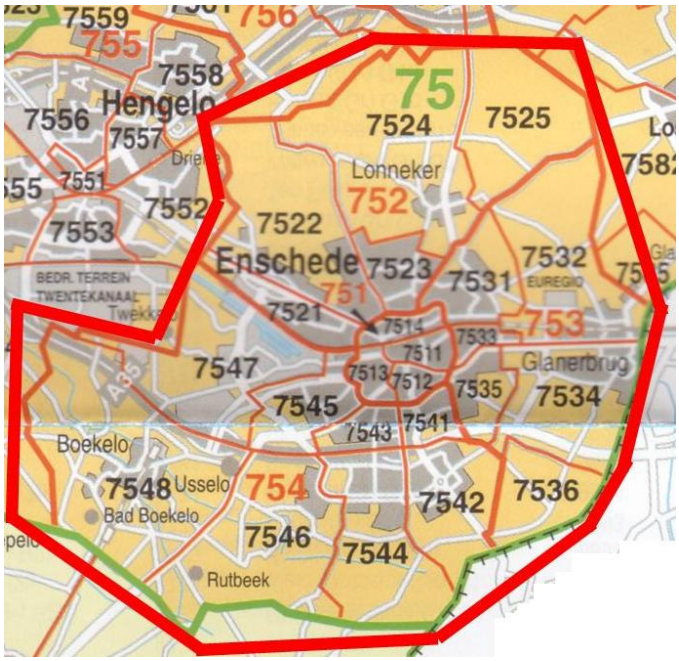


Figure 7.4 PC-4 areas in Enschede

7.2.7 SabiMIS

In Twente all busses are traced with Sabimos. This is a satellite-based navigation and monitoring system that guides busses and monitors where they are on what time. When this information is compared to the schedule it can be determined whether the bus has a delay or is in front of schedule. With that information the system can prioritise busses at intersections when they have a delay. With the data from the system a report is made about the operation each quarter: SabiMIS (in which MIS stands for Management Information System). In the report overall punctuality numbers are given and line specific punctuality numbers. In the report an overview is presented of the percentage of busses that are late or early is given (see figure 7.5 for an example). In the report there are different categories of one minute late or early. Examples are 1 to 2 minutes late and 3 to 4 minutes early. Services more than 6 minutes late or early are aggregated in the categories “more than 6 minutes early” and “more than 6 minutes late”. For the lines an overview is given of the punctuality on 2 terminal stops of the line and of the punctuality at all stops on each direction.

Categorie	<-6	-6 tot -5	-5 tot -4	-4 tot -3	-3 tot -2	-2 tot -1	-1 tot 0	0 tot 1	1 tot 2	2 tot 3	3 tot 4	4 tot 5	5 tot 6	> 6
gehele periode	0.5 %	0.46 %	0.91 %	1.82 %	3.52 %	6.86 %	13.3 %	20.52 %	18.06 %	13 %	8.2 %	5.08 %	3 %	4.84 %

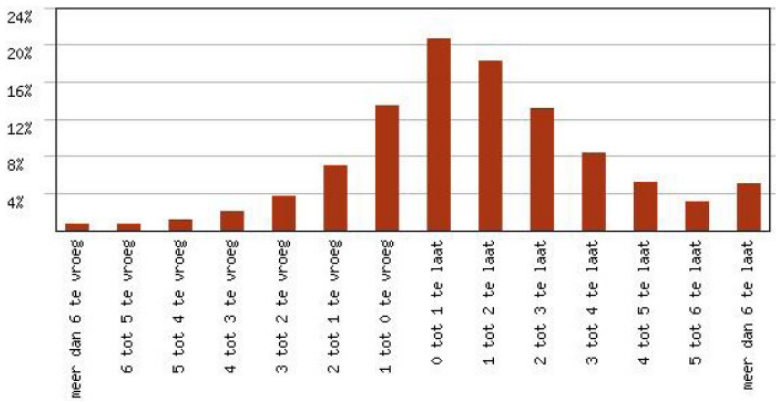


Figure 7.5 Punctuality at all stops in Q4 2006

7.2.8 MON

A lot of information can be found in the Dutch Mobility Research (MON) that is published yearly. In this research people have to report on their trips on a certain day. They report on origin, destination, time of departure and arrival, mode, distance and purpose. About 50.000 people fill in these surveys. For

this research differences between neighbourhoods are used to compare travel behaviour in those neighbourhoods. Therefore it is important to have a large number of surveys per neighbourhood.
In the research area about 200.000 inhabitants live. This is about 1,2% of all inhabitants of the Netherlands, which means that the expected number of people in the research area that filled in the survey is 600. When you consider that about 100 neighbourhoods are located in the research area, the average number of 6 people per neighbourhood can never be used to give a representative view of all people in those neighbourhoods. MON therefore unfortunately cannot be used.

7.3 Boardings per Area/Stop (per inhabitant)

The concession holder has multiple datasets about the number of passengers it transports. Which data to use are underpinned next.

PT-Chipcard
The number of boardings per zone/stop is the dependent variable in the research. Because of the introduction of the “PT-chipcard” the number of boardings per stop is known. This is a system in which a chipcard is used to pay for Public Transport, similar to the Oyster card in London. This card however was introduced in the study area not long ago (start 2010) and also other ticket types are still valid. Many people therefore will still use other cards, which makes that this source does not cover all boardings. The data therefore should not be used.

SLO
Another aggregated source of boarding numbers are the standard line researches (Dutch: SLO). For eight of the lines in the research area (2, 9, 20, 53, 59, 64, 73, 74) such an SLO was undertaken in 2007. The problem is that this is only about one third of all lines in the research area. The regional/rural lines are reasonably well covered, but only one urban and one inter urban line are included in the researches. The urban line is also changed between 2007 and 2008 which means that halve of the data cannot be used. Urban bus travel therefore cannot be modelled properly when these data are considered solely.

NVS-Counts
More disaggregated sources of boarding numbers are the NVS counts of the transportation companies. For these NVS counts all people are counted by the bus drivers. This is therefore a better representation of the true number of passengers/boardings. Another positive point is that the NVS Counts are done for all lines and not for a selection of lines in the research area.
A disadvantage of the NVS counts is that the number of boarders is not known for all stops but only for a group of stops. It is therefore needed to make aggregated zones when these numbers are used as the source for the number of boardings. However, not the number of boarders is known, but only the net number of boardings. This is because the people that deboarded before a counting stop are not counted.

When the SLO data and the NVS data are compared (figure 7.6) it appears that there is no systematic difference between the two. There however is a reasonable correlation between the two measures. In the figure for the SLO data only the number of boardings is used for the NVS counts deboarders are not counted. NVS counts therefore in theory should be below the SLO counts. In the figure it however can be seen that the trendline of the data has a formula that indicates that the number of boardings according to the NVS counts is higher than the number of boarders according to the SLO numbers (because it is above the line y=x instead of below). This all indicates that it is not possible to compute a reliable conversion factor for converting NVS data to SLO data. Because of the high correlation it however is possible to make some adjustments with the SLO data because no better information may be present in some occasions.

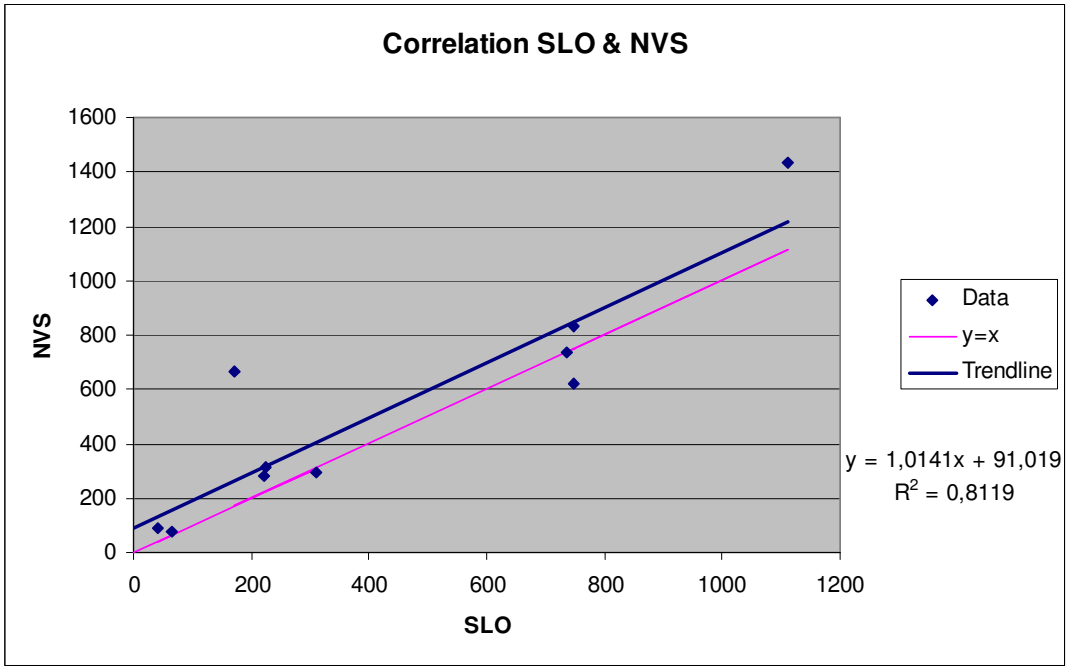


Figure 7.6 x-y plot of the number of boardings in SLO and NVS data

Using PT-Chipcard to disaggregate NVS counts?
The boardings in 2010 with the PT-Chipcard can be used to convert the NVS counts to boardings per stop. The relative number of boardings can then be used to subdivide the boardings in a zone to the boardings per stop in that zone. However, the current PT-chipcard use can have a bias because certain people might use the PT-chipcard less than others. This can be seen on line 1, where the number of boardings with PT-chipcard is relatively high till the first NVS counting-stop and relatively lower till the second counting stop. The reason for this is that the “neighbourhood” before the first counting-stop is the University of Twente area while the other is a “normal” neighbourhood (called Twekkelerveld). The difference between the two areas is the type of people that enter the bus. At the university mostly students enter the bus who already have and have to use the PT-chipcard. Other people, who mostly enter the bus in Twekkelerveld, in many cases do not own or use the Chipcard at the date of the measurement. Another interesting feature of the data is that for some lines (e.g. line 9) the number of boarders is lower than the number of deboarders according tot the PT-chipcard numbers. This is however technically impossible because the first time someone chips should always be seen as a boarding (because boarding is always done before deboarding). Altogether these findings suggest that it is better to not convert the NVS counts to boardings per stop using the PT-chipcard.

Conclusion
It is shown that it is best to take the NVS counts as the source for the number of boardings and make use of larger zones rather than stops as research objects (figure 7.7). When the PT-chipcard data are more reliable and cover all travellers, so when the card is at least the only valid travel ticket, a more detailed research can be done.

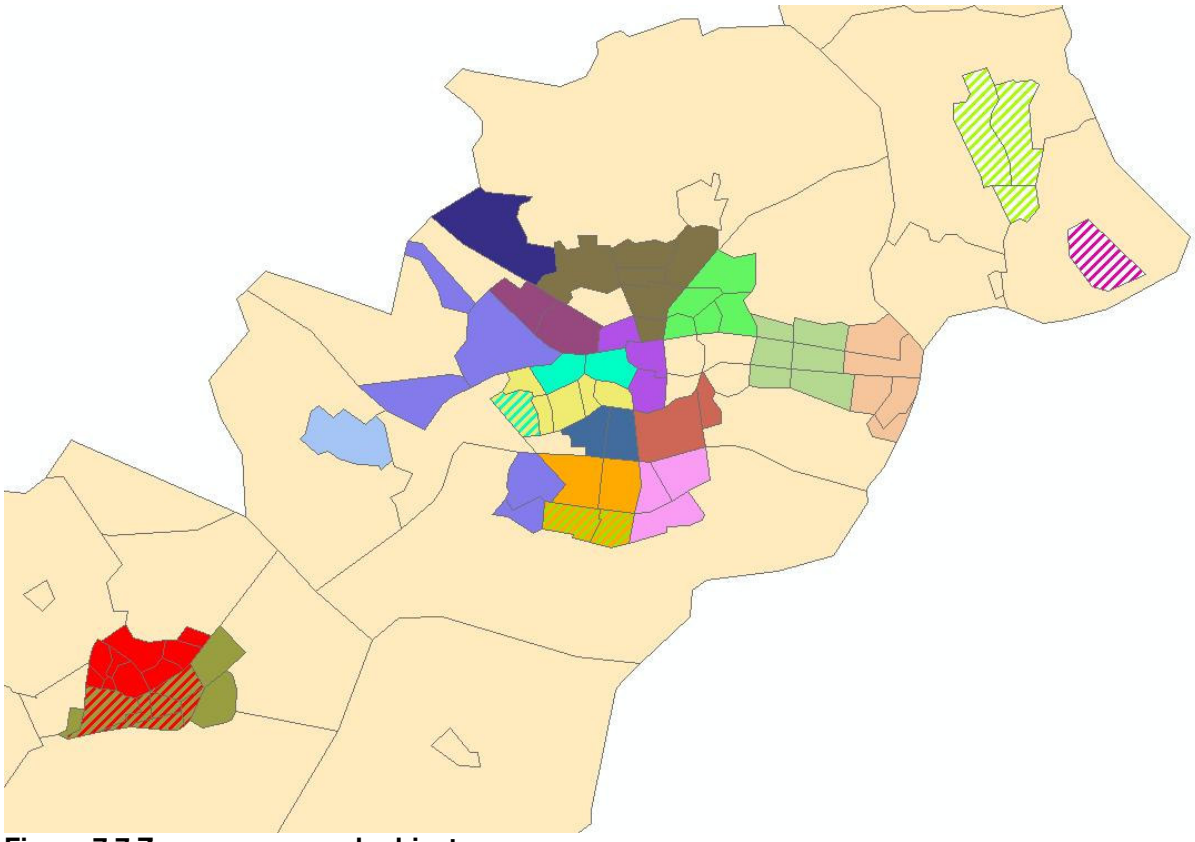


Figure 7.7 Zones as research objects

The boarding data that are used are of 2008, and, to be more precise, of November 2008. Because most lines run through one major trip attraction point (Enschede or Hengelo) it is assumed that all people that board are heading to the centre of one of the two cities. For the NVS counts only the number of passengers at a certain moment (stop) is counted. Therefore only the net number of boardings between two stops can be determined based on the counts. Because of the assumption that all people are travelling towards the city centre, the net number of boardings is equal to the total number of boardings. A problem with this assumption arises with busline 1 from the University to the city centre of Enschede at the stop near the second train station of Enschede, Drienerlo. There also a lot of people (mostly students, because all previous stops are on the University) deboard the bus to transfer to the train and travel further to their destination. To correct for this the number of boardings at the zone where this bus stop/train station is located is enlarged with the number of deboardings by PT-chipcard in April 2010, which is 102 (Connexxion, 2010), which is currently the best guess of the number of (de)boardings. Because students are already forced to use the PT-chipcard from this month[7] and because nearly only students travel at this part of the line this should be quite reliable. Although this number is not 100% reliable because not everybody uses the PT-Chipcard already it is better than not taking into account the number of deboarders at Drienerlo at all.
In Haaksbergen there is also a slight problem because no NVS counts that can be attributed to this zone of lines 73 and 74. This is corrected with help of the SLO for these lines that are fortunately available. The total number of boarders is first compared to the number of people aboard at a stop where NVS counts took place and the number of boarders per stop is than adjusted with the fraction of these numbers. Then the number of boardings is known. This is one of the exceptions mentioned earlier where it is not possible to use direct NVS counts, but the SLO data are used.

Boardings per inhabitant
For the research a multiple linear regression model is estimated of the demand for bus transport. When the absolute number of boardings is used as the dependent variable some problems arise with this linearity. The most important problem is that a wide variety of boardings per zone are present. In some zones only 34 boardings per day take place whereas in other zones as much as 931 boardings take place. In the city centre the number of boardings is even more extreme (above 13.000).

[7] http://www.9292ov.nl/9292ov1768.asp

When some variables, such as car ownership per inhabitant or the percentage of students, are included in the linear model an increase of the variables always leads to an absolute change in the number of boardings for each zone. The relative influence however is very different in a zone with 34 boardings than in a zone with nearly thirty times as many boardings.
To correct for these problems the number of boardings per inhabitant is used as the dependent variable. The variable values are in the range of 0.007 to 0.288 boardings per inhabitant (except the city centre). This range is still quite wide, but this mainly depends on two values that are far from the other values. One is the value for the university (0.288), which is the effect of large numbers of travellers (especially students) and the low number of inhabitants. The second highest value is 0.109. For some other variables, which could have been used in a model for the absolute number of boardings, this means that the values also have to be divided by the number of inhabitants. In figure 7.8 the number of boardings in each zone is graphically displayed. It can be seen that the number of boardings per inhabitant is quite high in Haaksbergen and Overdinkel and in some outskirts of Enschede.

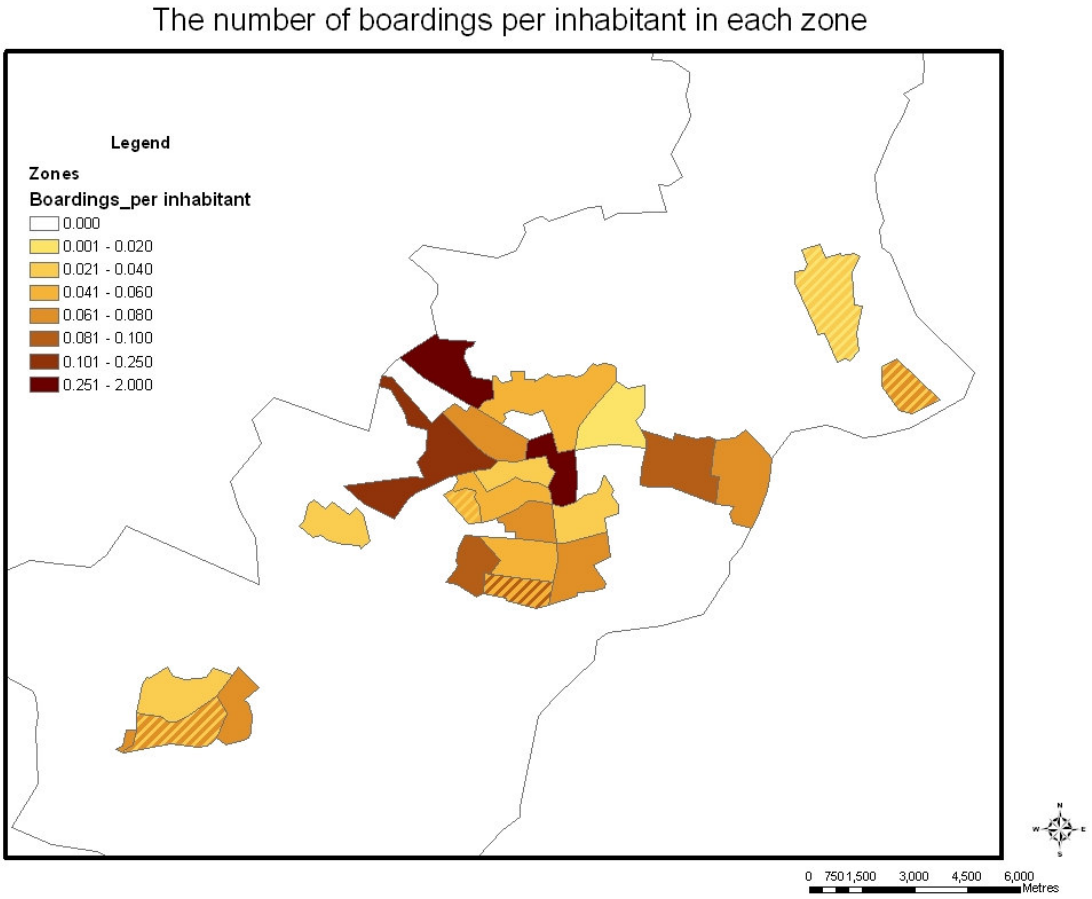


Figure 7.8 Boardings per inhabitant

7.4 Socio economic

Four socio economic variables are included in the research: car ownership, business car ownership, income and the percentage of students.

7.4.1 (Business) Car ownership

The car ownership data that are used come from the Neighbourhood map of Dutch Statistics. These data, on a sub-PC4 level scale, are available for the years 2004 to 2008. The measure of car ownership is the number of cars owned per inhabitant.
Business cars are defined by De Jong, Fox, Pieters, Daly & Smith (2004) as “cars of self-employed persons who registered the car in the name of their firm or cars provided by employers to their employees, either owned by the company or leased”. In the Neighbourhood map however business cars are only minibuses, lorry’s, busses and special purpose vehicles. Information on lease cars is not

available. A problem arises because business cars can be put on name of both the company and the car driver. Therefore the data on business cars are not very reliable (and not complete).
Because no data that are more reliable are available for other business can be used as determinant. This is answered in the regression phase.

Aizcorbe & Starr-McCluer (1997) have shown that owning a lease car is highly correlated to (household) income. Therefore the differences in the demand for bus transport because of the "ownership" of lease cars works through in the differences in bus demand because of different income levels. The (relative) number of lease vehicles therefore is not included in the research.

7.4.2 Income

The income in the different neighbourhoods and villages is registered by the municipalities and by Dutch Statistics. The scale differs for the different sources. DS for instance has data available on their website of 2004 and the municipality of Enschede has data about 2007 in their Neighbourhood monitor. In this monitor for all the zones the average income per head/household/worker is included. The data of Dutch Statistics, which is the best data for Haaksbergen and Losser, only are about standardised household income. Therefore standardised household income is used as the variable in the research. To compensate for the three years in between the measurements of the whole research area and Enschede alone the values for the neighbourhoods in Haaksbergen and Losser are increased by the average increase of standardised household income in Enschede (about 12.0%). At a glance an increase of 12% in three years seems to be very high. The Economy of the Netherlands increased 9,3% (2.0, 3.4 and 3.6%) in total in the same period. In the province the research area lies in this is even 9.9% (1.8, 3.0 and 4.8%). Together with an increased participation on the labour market it makes this 12% increase not improbable.
Because some zones have very few inhabitants there are no data for them. For these areas the income is assumed to be equal to the average of the income of the whole municipality it lies in. Because not many people live here the bus demand potential is also not high which makes that the errors made by doing so do not propagate severely in the final model. Together with the unreliability of the data (see 7.2.4) it should be kept in mind that these data are not ideal. The income in the research objects is the weighted average of the income with respect to the number of inhabitants per neighbourhoods.

7.4.3 The percentage of students

The number of students per PC-4 area comes from the registering agency OCW DUO. The data are of the 2009-2010 curriculum, which means that they do not align perfectly with the study year 2008. Because the registering agency is the source of the data the reliability is high.
For study year 2008 the numbers of students are different than the numbers that were supplied by DUO. To check whether the data can be used the total number of students per zone is compared to the 2008 numbers of inhabitants in the different zones. It appears that there is a large correlation between the percentage of students in 2010 and the percentage of people aged 15-24 in 2008 (see figure 7.9). In this figure the area in which the campus of the university is located is excluded because of the otherwise disproportionally high importance of this value, which is an outlier because many students live on the campus.

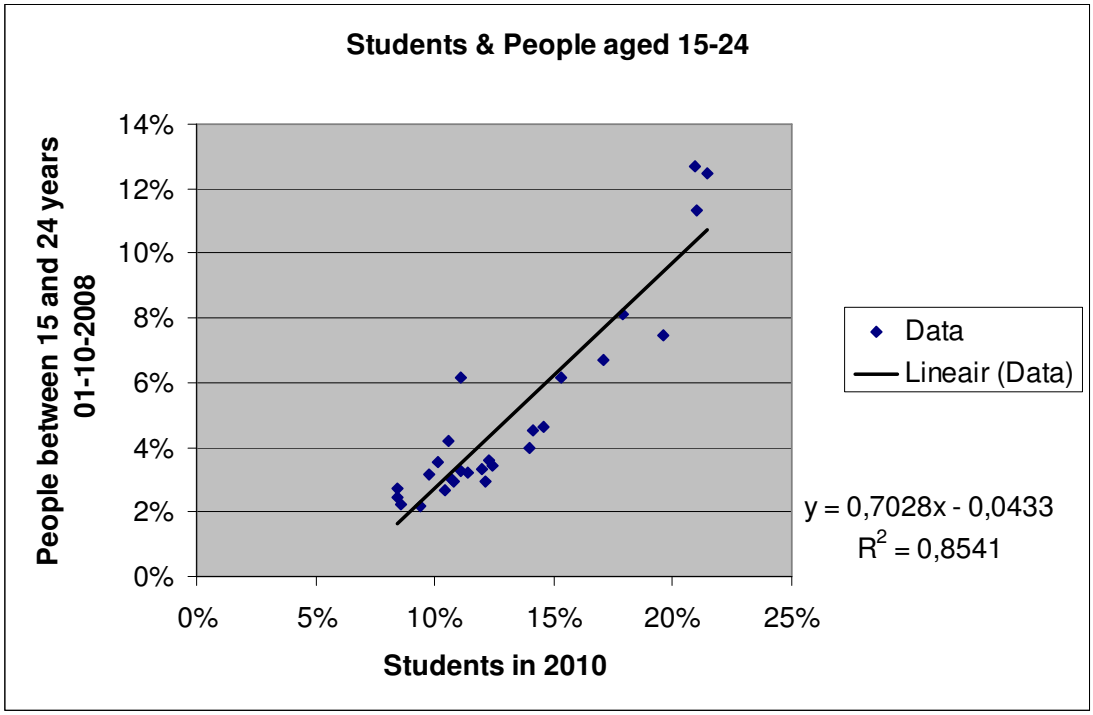


Figure 7.9 Compare Students 2010 to Age Groups 2008

A large correlation between the two variables means that both variables can be used to as a measure for the numbers of students. The number of students from OCW DUO are however more precise because all people included really are students. Another advantage of the DUO-data is that a distinction is made between high- and middle-level education and between part- and full-time students. Because of that, some students can be excluded because they do not get a Student Public Transport card. This card can be used for free or discounted travelling and is only for students/scholars that are above 18 years and enjoy full-time education[8]. Therefore scholars of part-time lower education are not included in the number of students. Especially on middle-level (mbo) education, there are also scholars under 18 years. No Student PT-card is available to them, which means that they have to pay normal prices for travelling by Public Transport. On average an mbo study takes 4 years and people start such a study when they are 16 years. About one half of the total time the study is followed the scholars therefore have a Student-PT card. Therefore the number of mbo students is divided by two.
An advantage of the number of people aged 15-24 is that the data are about the study year (2008) and that the spatial scale level is more detailed (neighbourhoods instead of PC-4 zones). The high correlation between the two variables indicates that both variables can be used as a measure for the number of students in a neighbourhood.

Conclusion
The 2010 data from OCW DUO are most precise, but the 2008 data from the neighbourhood map are more detailed (on a neighbourhood level) and are about the study year itself. To check what the consequences are of using the different measures for the number of students both variables are used in the regression and the differences between the outcomes are described.

7.5 Built environment

The built environment variables that are included in the research are the number of inhabitants and jobs, the number of college enrolments in an origin zone, the accessibility of destinations and the spatial density.

[8] http://www.ib-groep.nl/particulieren/studiefinanciering/sfmbo/voorwaarden.asp

7.5.1 Inhabitants & Jobs

The number of inhabitants and jobs comes from the Regional Traffic Model Twente (RVM Twente) of 2008 which means that the data in it are based on 2008 numbers. Because the municipalities have cooperated in the development of the model and the municipalities register their inhabitants the data in the model are very reliable.
The division of the research area into zones also exists in the neighbourhood map of Dutch Statistics. The zones in the RVM are however a lot smaller than the zones in the neighbourhood map. The GIS programme (ArcGIS) can be used to unite (see figure 7.10 for an example of the union of two layers) the two different geometries into one. The information in the different layers is also united, which means that both the area numbers of the regional model are known as the information from the neighbourhood map. With this information the number of jobs and inhabitants per neighbourhood can be calculated.

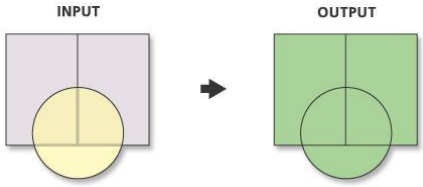


Figure 7.10 Union example

Because of the resolution of the data about the number of boardings, the RVM zones have to be aggregated to neighbourhoods. When one RVM zone lies in more than one neighbourhood the number of jobs and inhabitants are divided between the two based on the relative area of the different parts of the RVM zone. A threshold to this is that at least 25% of the RVM-zone should be outside the neighbourhood it belongs to most. When one RVM zone lies for 25% in one neighbourhood and for the other 75% in another neighbourhood the first one gets 25% of the inhabitants and jobs and the other one gets the rest. This threshold is used to exclude the problem of borders that are actually meant to represent the same border between neighbourhoods, but do not do so because of errors in spatial referencing (like in figure 7.11).

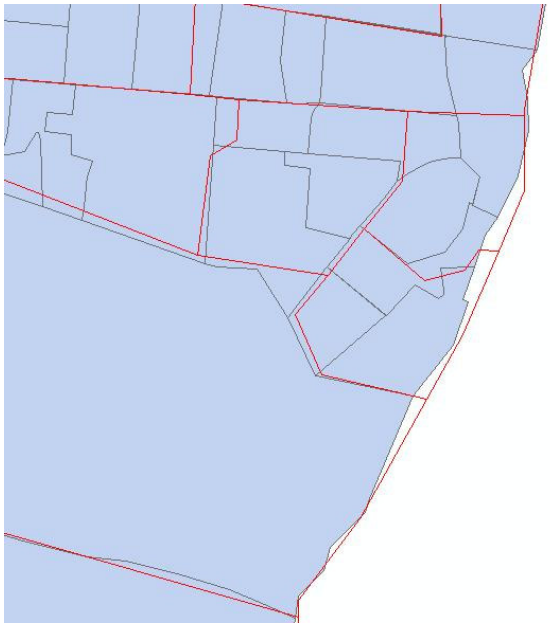


Figure 7.11 Non-identical borders near Glanerbrug

7.5.2 College enrolments

In the study area different education institutes are located. Primary and secondary education institutes are not included in the research because scholars of these institutes don't travel for free or get a discount when they travel by bus. The institutes that are included are: University of Twente, Saxion Hogescholen, ROC of Twente and ArtEZ.
The university has an overview of the number of students that are enrolled on their education programmes. For every year (1964 to 2008) the number of students can be found on their website[9]. From 2002 on a more detailed overview is present which makes a distinction between bachelor, Master and Premaster students. The University is located on a campus in the North-West of Enschede and consists of several buildings. On and close to the campus

Figure 7.12 Bus stops near the

[9] http://www.utwente.nl/feitenencijfers/onderwijs/totaal/inschrijvingen.doc/

there are 13 bus stops (figure 7.12), of which 5 (red without dot) can be seen as stops to service the education part of the campus (purple areas) and 8 to service the living part of the campus (blue with dot). The reliability of the number of students is high because it comes directly from the university.
Saxion Hogescholen is located in the centre of Enschede, near the central train station. On Saxion's annual report website[10] the number of students in the years 2007 and 2008 can be found for their location in Enschede subdivided into full-time, part-time and dual students. Because the data come from the annual report of Saxion the reliability is high.
The ROC of Twente is a middle-level (mbo) education institute. It has two large locations in Hengelo and Almelo, bus also has some smaller locations in Enschede. The buildings are located on five different locations in the city. For all these locations the number of students in the years 2001-2009 (on 01-10) are known. The data come from the ROC itself and are therefore very reliable. The data also distinguish between part-time and full-time education. Part-time in most cases means that students work 4 days and go to school one day a week. Part-time students do not get a Student-PT card. Therefore only full time students are included in the number of enrolments. Again, because mbo students only get a PT-card for free when they are 18 years or older, the total number of mbo enrolments are divided by two.
ArtEZ is the school of arts of the East of the Netherlands. In Enschede it has two locations, one on the campus of the University and one in the Muziekkwartier, near the central train station. From its annual reports the number of students on both locations for the years 2002 to 2008 can be found. Again reliability is high because the source is the educational institute itself.

7.5.3 Accessibility of destinations

The accessibility of destinations can be represented by different variables. These are the relative accessibility and the (relative) travel time towards the city centre.

Relative accessibility
The accessibility of destinations of the bus compared to other modes in a certain neighbourhood can be represented by the relative number of jobs and inhabitants that can be accessed by the different modes (PT in relation to car and bicycle). The web tool Accessibility map Netherlands[11] gives insight in the accessibility from a certain origin on PC-4 level. To do so the travel time isochrones are computed. Travel time isochrones are lines that are accessible with the same travel time. All people within a certain isochrone are all accessible within the travel time that belongs to that isochrone.
The Regional Traffic Model (RVM Twente) is developed by the same consultancy firm as the Accessibility map. The programme that the models work on also is the same (OmniTRANS). With the regional model therefore also travel time isochrones can be computed to determine how many jobs and inhabitants are accessible from a zone in the regional model. The most actual model is the 2008 model. The reliability of the location of the jobs and inhabitants in the zones is again very high, because the municipalities have cooperated during the model development. The travel times are somewhat less reliable because the traffic numbers come from a model and not from the real world. Because obtaining the data from the real world directly is too complex the available information is the best possible information. In appendix C a more detailed description of the determination of the relative accessibility of the different zones is given.
In formula form the relative accessibility is: $\frac{\#Jobs_{Car}}{\#Jobs_{PT}}$ and $\frac{\#Jobs_{Bicycle}}{\#Jobs_{PT}}$

(Relative) travel time towards the city centre
A disadvantage of the relative accessibility is that the train plays an important part in the accessibility by public transport. Therefore another measure is used that only looks at bus, car and bicycle travel. Earlier it was assumed that all boarders travel towards the city centre of Enschede (or Hengelo and Oldenzaal, for some lines). The relative travel time towards this centre therefore can be important for the relative number of bus users. The travel time by bus, car and bicycle therefore is determined using the schedule and Google Maps[12]. The travel time by bicycle is represented by the travel distance walking times 15 kilometres per hour, which is assumed to be the average velocity of bicyclists. The variables that are included in the research are the travel time towards the centre by bus and the relative travel time by car and bicycle.

[10] http://www.jaarverslagsaxion.nl/
[11] http://www.bereikbaarheidskaart.nl/
[12] http://maps.google.nl

The relative travel time by car is: $\frac{TT_{Car}}{TT_{Bus}}$, and by bicycle: $\frac{TT_{Bicycle}}{TT_{Bus}}$

A larger value of the relative travel time indicates that the travel time by either car or bicycle is relatively larger than the travel time by bus.

7.5.4 Spatial density

The spatial density of a neighbourhood is represented by three measures, the population density, the job density and the address density.

Population and job density
The population and job density is the number of inhabitants/jobs in a certain neighbourhood (from the RVM) divided by the area of the neighbourhood, that can be determined using ArcGIS.

Address density (OAD)
Dutch Statistics uses the address density in the neighbourhood (Dutch: Omgevingsadressendichtheid, OAD) as a representative for the level of urbanisation. This OAD is determined by counting the number of addresses that are located within a circle of one kilometre around a certain address. When the OAD of a certain neighbourhood is equal to 1000, on average, 1000 addresses are present within a circle of one kilometre of all addresses. This value of the OAD is used as the last possible measure of the spatial density.

7.6 Transportation network

Also four transportation network variables are included in the research. They are: frequency, punctuality, train station dummy and distance to stop.

7.6.1 Frequency

The different lines through the research area have different frequencies. Some lines are HOV lines (High quality Public Transport), which means that high frequent, fast and reliable transport is offered, while others are only used for low density passenger flows and are operated once every hour. The timetables of all the lines are written down by the bus operator in a timetable book which can be used to determine the frequency on a working day. For all years from 2002 these books exist. Because the bus operating company publishes these books the timetable information is very reliable. This however does not include whether or not the busses are on time or driven at all.

7.6.2 Punctuality

Punctuality is dependent on the possible delay on a certain line. By building designated infrastructure, such as free bus lanes and streets, and prioritising busses at intersections the punctuality (and the travel time) can be improved. An advantage of this variable hence is that it can be influenced by governmental organisations.
The punctuality numbers of SabiMIS are used. The Sabimos system and the corresponding SabiMIS report are not perfect, because in some cases it does not work and also sometimes strange outcomes are given (for instance: many busses leaving more than 6 minutes early from the first stop of a line). Especially the punctuality at the terminal stations of lines seems to be very strange on multiple occasions. Therefore not the stop-specific punctuality numbers, but the line-specific punctuality numbers are used. Disadvantage is that the punctuality on all stops is assumed to be equal which is not true. The reliability of the data however is much higher than when the punctuality of the different stops is estimated. The punctuality of a line is defined as the percentage of stops that is between one minute early and three minutes late. In this way at least a good overview of the differences between the different lines can be given.

7.6.3 Train Station

Because busses are used as a complement to trains very often the number of boardings near the central train station of Enschede is relatively high. To include this in the model an extra variable is included that has a value of 1 when it is the stop/zone near the train station and 0 otherwise. This type of variables is called dummy variables. Because the location of the train station is known the allocation of the station to a neighbourhood is very easy.

7.6.4 Distance to stop

The distance to a stop is highly dependent on the location of the stops in the neighbourhoods. The average distance can be determined with a GIS programme that has the borders of the neighbourhoods and the location of the stops in it. The zonal borders are available in the neighbourhood map and the coordinates of the stops are available at the transportation companies. The stop coordinates are from 2009. Because stops do not change very often the locations are the same as the years before. Only in some cases lines have changed and stops have been entered or exited service. Because the information comes form the transportation companies it should be very reliable.
Stops are most of the times on two sides of the road because lines go in two directions. These stops are aggregated to one because otherwise strange influence areas (one of the stop in one direction and for the stop in the other direction) of stops are computed and because a trip to and from a certain location is usually undertaken with the same mode and line. The aggregation is done based on the coordinates. The stop that represents the two stops lies in between the other two (see figure 7.13). In the figure the two stops are S_1 and S_2 and the aggregated location of the stops is S_R.

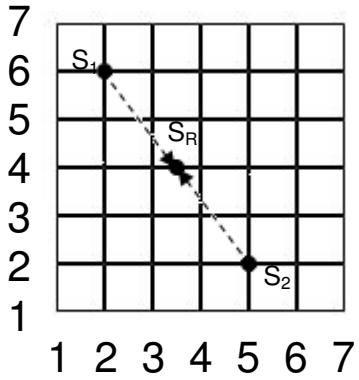


Figure 7.13 Determining the coordinates of the representative of the stop

Aggregating two stops to one has some implications. When you assume that the catchment area of a stop depends on the number of people living within a certain distance from the stop a problem can occur (see figure 7.14). The catchment area of the left stop (area 1) is equal to A+B while the catchment area of the other stop is equal to B+C. After aggregating the stops (area 3 is the assumed catchment area of the aggregated stop) the catchment area is said to be as large as D, but it can be that only people that live in the catchment area of both stops consider the distance to the stop short enough (equal to B). The same accounts for the distance from a house towards the nearest bus stop. It is possible that people consider the distance towards a stop to be the distance to the furthest stop, while the distance towards the aggregated stop is used in the research.
On average, and especially in urban environments the distance between two stops is not very large. In most cases two stops are on the other side of the road. Therefore areas A and C are not that large which means that B does not differ that much from D. To keep the computability of the model easy the aggregation of the stops stays intact.

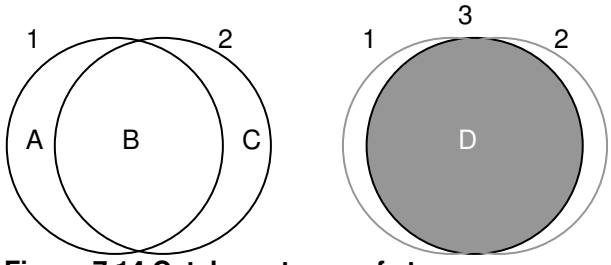


Figure 7.14 Catchment area of stops

When the locations of the stops are known it is possible to determine for all the neighbourhoods (most detailed scale used by Dutch Statistics) what the average distance is to a stop with help of a GIS programme. The distance to a stop is the straight line (Euclidian) distance between a certain location and the nearest bus stop.

Or in a formula: $DtS_i = \sqrt{(x_i - x_s)^2 + (y_i - y_s)^2}$

Where:

DtS_i = Euclidian distance to stop of location i

x_i = x-coordinate of location i

x_s = x-coordinate of the nearest stop

y_i = y-coordinate of location i

y_s = y-coordinate of the nearest stop

The process of the determination of the average distance to a stop with the GIS-programme can be found in appendix D. The outcomes are shown in figure 7.15. In this figure lighter areas have shorter average distances towards stops than dark areas.

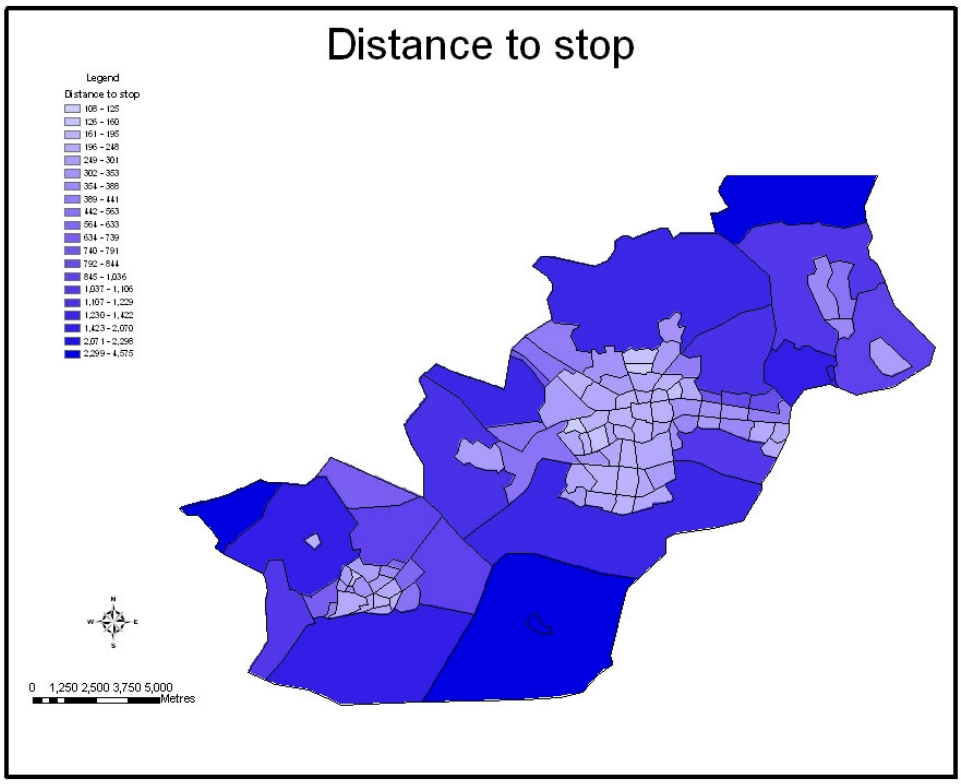


Figure 7.15 Average distance to a stop per neighbourhood

7.7 Conclusion

For all variables data are available. Not all datasets are however available for all years. For 2008 all variables are known except the income of people or households. Because the income in 2007 is known and income does not change that much in one year the 2007 data can be used to represent the 2008 values of the average income in a neighbourhood. 2008 therefore is the year that is used for the research.

To check whether the model that is determined for the number of boardings is just a good representation of the 2008 data or could also be used for other years, another year is used for the validation of the model. Because it is interesting to see what numbers are estimated by the model in the previous concession period the validation year should be before 2006 (in December 2005 the concession started). Because most data are available for 2004 this year is used as validation year.

8 Correlation

Before the first regression is done it should be studied whether the dependent variable correlates with the different independent variables and whether the different independent variables correlate with each other. When independent variables do not correlate with the dependent variable including it in the model probably does not have positive effects. When the independent variables correlate to high extend (multicollinearity) both variables can be detected as non-significant while including only one can indicate a significant relation with the dependent variable. In this chapter the correlation in the dataset is located and, most important, explained.

8.1 Dependent variable

Pearson's correlation coefficients are determined for the dependent variable (the number of boardings per inhabitant per zone) with the different independent variables. First the correlation with all zones included except the city centre is discussed. After that the correlation when leaving two more zones, the university and the harbour, out is discussed.

8.1.1 All zones except the city centre

To give an idea about the relation between the sole independent variables and the dependent variable the correlation coefficients can be determined. Variables that do not relate to the number of boardings per inhabitant are not likely to explain the variance and hence should not be used in first instance. The correlation coefficients of the relation between the number of boardings per inhabitant in a zone and the different independent variables are presented in table 8.1. The coefficients are determined using the research object spatial scale (line specific zones). The city centre of Enschede is not taken into account as a zone because the number of boardings per inhabitant is very high in this zone because most people board here at their way home. The number of boardings in this zone cannot be translated to a number of boardings per inhabitant because the boardings in many cases do not belong to people living in the centre.

Variable	Correlation	Variable	Correlation
Car ownership	0.026	Travel time to city centre by bus	0.037
Business car ownership	0.148	Relative Travel Time car	0.080
Income	0.265	Relative Travel Time bicycle	-0.185
% of students	0.754	Population density	-0,283
% of people aged 15-24	0.829	Job density	0,244
Inhabitants	-0.361	Address density (OAD)	-0,163
Jobs / Inhabitant	0.187	Frequency	0.406
College enrolments / Inh.	0.829	Punctuality	0.116
Relative car accessibility	-0.310	Train Station	unknown[13]
Relative bicycle accessibility	0.172	Distance to stop	0.226

Table 8.1 Correlation with the number of boardings per inhabitant per zone

Not all independent variables are correlated with the number of boardings per inhabitant. The highest (absolute) correlation coefficient is with the percentage of people aged 15-24 and college enrollments (0.829).

Remarkable values
In the table above also some values are present that conflict with common sense or literature.

Car ownership & income
(Business) car ownership and income, both variables related to the wealth of people, are all positively related to the number of boardings per inhabitant, although the coefficients are not really high. People having a higher income and owning more cars however are thought to use the car more and hence do not take the bus that often. A reason for this unexpected sign can be that car ownership is especially high in business areas where few people live and people travel from these areas back home. Other

[13] Because the zone with the train station is not included in the determination of the correlation no coefficient could be determined

people than the people living in the area hence board in these stops which makes the number of boardings per inhabitant higher than average.
The harbour area has high numbers of boardings per inhabitant and also high numbers of car ownership and income. Especially the car ownership numbers are extremely high in this zone. This can lead to the unexpected sign for the correlation between the independent variable and the number of boardings. When the zone is excluded the correlation coefficients become -0.312 (car ownership), 0.011 (business car ownership) and -0.339 (income). For two of the three variables the sign than becomes as expected. Business car ownership is really not correlated to the number of boardings, which confirms the questions posed earlier on the reliability of this variable.

Relative bicycle accessibility
The relative bicycle accessibility has a positive correlation with the number of boardings per inhabitant. Expected however is that in areas with a good bicycle accessibility less use is made of busses. A possible explanation can be that the number of boardings is higher in urban areas where bicycle accessibility is also highest. Another possible explanation can be that it is a problem that the accessibility of both bus and train are used in this measure and not of bus alone. This indicates that it might be better to use another measure for the accessibility of destinations of the bus compared to other modes.

(Relative) travel time towards the city centre
The travel time to the city centre by bus and the relative travel time by car do not seem to be related to the number of boardings per inhabitant. The low correlation coefficients (0.037 and 0.080) indicate that no clear relation is present. One would however expect that when the relative travel time by car is high more people use the bus and that longer bus trips lead to fewer boardings. A clear reason for the lack of correlation cannot be found, but it might be a good idea to look at the correlation without two extreme zones. These zones, the university and the harbour, have high values for some variables which can lead to large deviations. Another explanation can be that the car is not a real alternative for travel by bus. In that case the real choice of potential bus travellers is between bicycle and bus.
The relative travel time by bicycle is negatively related. The larger the travel time by bicycle is the lower the number of boardings per inhabitant. This is the opposite of the expectations, but might be explained by the fact that long travel times by bicycle encourage car use instead of bus use.

Population and address density
Both the population and the address density show a negative correlation to the number of boardings per inhabitant. The lower the population or address density is, the lower the number of boardings per inhabitant hence is. This is not in accordance with different studies (Balcombe et al , 2004, Pauley et al, 2006) that state that more trips are made by bus by people living in densely built areas. In the research area however it is the case that the number of boardings is relatively higher in neighbourhoods far from the centre and in the rural municipalities. These zones however are not that densely built. In the more densely built areas near the centre of Enschede the number of boardings is lower because bicycle accessibility is better there and slow modes (bicycle, walking) can be used as access mode towards the train station. Together this seems to be a plausible explanation of the negative sign of the correlation coefficient between these variables and the number of boardings per inhabitant. The fact that the sign is different in the research area than according to literature is no reason for excluding the variables from the research as they do explain a difference in the demand for bus transport.

Distance to stop
The distance to a stop is positively correlated to the number of boardings. This means that when the distance is higher the number of boarding is also higher. This contradicts expectations because it seems logical that the closer stops are the higher the number of boardings. A reason for this finding can be that the distance in the study area is never very large (in the range between 175 and 465 metres). This means that most people can reach stops in a small amount of time. The correlation is however not that high which could mean that together with other variables it does determine something of the number of boardings. The variable therefore stays included.

Rural and urban zones
In the research both rural and urban zones are included in the same dataset. It is however interesting to look at the differences in correlation for rural and urban zones. A disadvantage is that only 15 (urban) and 7 (rural) zones are included in the research. The correlations can be found in appendix E.

Because of the low number of data points it can be seen that the correlations do differ a lot. Also signs are different in many occasions. Another interesting thing is that most rural and urban correlations are more significant than the overall correlations. For the relative bicycle accessibility another strange thing happens. The correlation of both the urban and the rural zones is negative, but the correlation for the zones together is positive. The reason for that is the large difference in the relative bicycle accessibility. The rural zones most have values of between 0.1 and 0.2 while the urban zones have values of 0,4 and above (see figure 8.2). Together this hence leads to a positive instead of two negative relations.

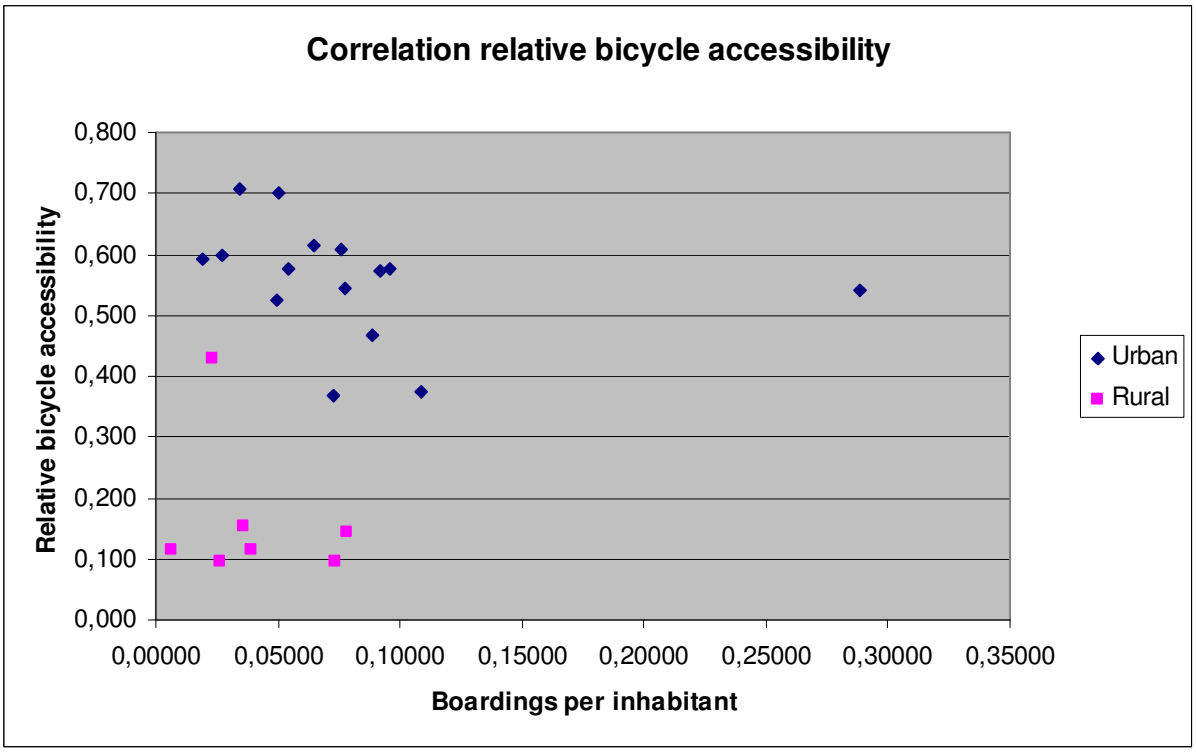


Figure 8.2 The difference in correlation between urban and rural zones

It is of course not beneficial for the research that this difference occurs. Making one model for urban and one for rural zones however is impossible because too little zones could be used for the regression. Therefore the rural and urban zones are used together.

8.1.2 Leaving the university and harbour zone out

This time the zones around the train station and the university are excluded because the number of boardings in those zones cannot clearly be assigned to the inhabitants in those zones. Also the harbour is excluded because that is a zone where only businesses are located which leads to unrealistic car ownership numbers.
The outcomes can be seen in table 8.2.

Variable	Correlation	Variable	Correlation
Car ownership	-0,312	Travel time to city centre by bus	0,126
Business car ownership	0,011	Relative Travel Time car	-0,114
Income	-0,339	Relative Travel Time bicycle	-0,217
% of students	-0,027	Population density	0,014
% of people aged 15-24	0,087	Job density	0,232
Inhabitants	-0,160	Address density (OAD)	0,144
Jobs / Inhabitant	0,169	Frequency	0,766
College enrolments / Inh.	-0,293	Punctuality	0,294
Relative car accessibility	-0,014	Train Station	-
Relative bicycle accessibility	0,198	Distance to stop	-0,066

Table 8.2 Correlation without the university and harbour zones

Remarkable values
What is interesting is that some coefficients have changed sign. This happened for the car ownership, income, the percentage of students, the number of college enrolments per inhabitant, the travel time to the city centre by bus, the relative travel time by bicycle, the population and address density and the distance to stop.
For car ownership, income, the relative travel time by bicycle, the population density, the address density and the distance to stop the sign now is what one would expect. It seems that the extreme value of the number of boardings per inhabitant in the university zone really influenced the correlation coefficient negatively for these zones.
For the percentage of students and the number of college enrolments per inhabitant the university however made that the sign of the correlation was as expected. Because the other zones do not have large differences in the percentage of students the value of the correlation coefficient of the percentage of students (and also the percentage of people aged 15-24) is very low. Because only one zone still has (a very small number of) college enrolments the correlation coefficients of the number of college enrolments and the number of boardings become meaningless.

Conclusions
With the lower number of zones included the number of boardings per inhabitant is only significantly correlated to the frequency of services (coefficient: 0.766). That these two variables are correlated could be expected because the number of boardings in a zone usually increases when the frequency is enlarged. The frequency therefore is the variable that really should be included in the model for the number of boardings (per inhabitant). This also corresponds to conclusions of Paulley et al. (2006) and Bresson et al. (2003), who show that frequency (or the number of vehicle kilometres) is an important determinant for the number of travellers.
The other variables do not show this clear correlation, but including them in the model could increase the amount of variance explained by the model (the goodness of fit). Interesting variables are those with high correlation coefficients that have the sign that could be expected. Examples are: car ownership (-0.312), income (-0.339), percentage of people aged 15-24 (0.087, not high, but good sign), the number of jobs per inhabitant (0.169), the relative travel time by bicycle (-0.217), the job density (0.232) and punctuality (0.294).

8.2 Independent variables

Now the correlations between the different independent variables are described. The correlation is determined for both neighbourhoods, as many data are also available on this level, and for zones, which are the research objects. Again the train variable is not included because only one zone has a train station.

8.2.1 Neighbourhoods

For the mutual correlation of independent variables it is important to determine a threshold for the correlation coefficient. When the coefficient is larger than this threshold one of the two variables should in principle be excluded from the research. According to Kuby et al. (2004) the danger value of Pearson's correlation coefficient is (+/-) 0.7 which means that the threshold should be that value. To also take into account slightly correlated variables the cases in which the value of the correlation coefficient is higher than (+/-) 0.5 also are described and, if possible, explained.
In appendix F an overview of the correlation between the different independent variables on the neighbourhood spatial resolution is given. The correlation is determined using all neighbourhoods that are serviced by a bus line in the research area, which makes the number of observations higher and the value of the correlation coefficient more reliable. Correlation coefficients between -0.5 and 0.5 are marked green in the appendix which means that these correlations are not in the danger zone and will not be considered separately. Correlations between -0.7 and -0.5 or 0.5 and 0.7 are marked orange which means that they are not in the danger zone but will be considered. Red marked correlation coefficients are lower than -0.7 or higher than 0.7 and are in the danger zone. When the coefficients are in the danger zone the two variables, in principle, should not be included in the same model.

Non-danger zone
The correlation coefficients are discussed from top left in appendix F to bottom right.

Percentage of people aged 15-24 and percentage of students
As could be expected the number of students and the number of people aged 15 to 24 are correlated. It is more interesting to discuss why the correlation is not closer to 1 than 0.676. The reason for this most likely is the level of detail of the different variables. The number (and percentage) of students is known at a PC-4 level while the percentage of people aged 15-24 are known at a neighbourhood level. In most PC-4 areas usually more than one neighbourhood are located which means that the number of students is more or less related to the average of the percentage of people aged 15-24. In figure 8.3 both the average and the individual X-Y plots can be seen. In the graph the blue diamonds represent neighbourhoods and the pink squares represent the PC 4 areas. It can be seen that the zone with 25% students really influences the correlation negatively because the PC-4 zone can be subdivided in three neighbourhoods that have different percentages of people aged 15 to 24. The difference is that large because one of the neighbourhoods is the campus of the university where many students live. Because the level of detail is higher for the number of people aged 15 to 24 it might turn out that the number of students can best be represented by the number of people aged 15 to 24. To determine whether that is correct both variables should be included in the regression.

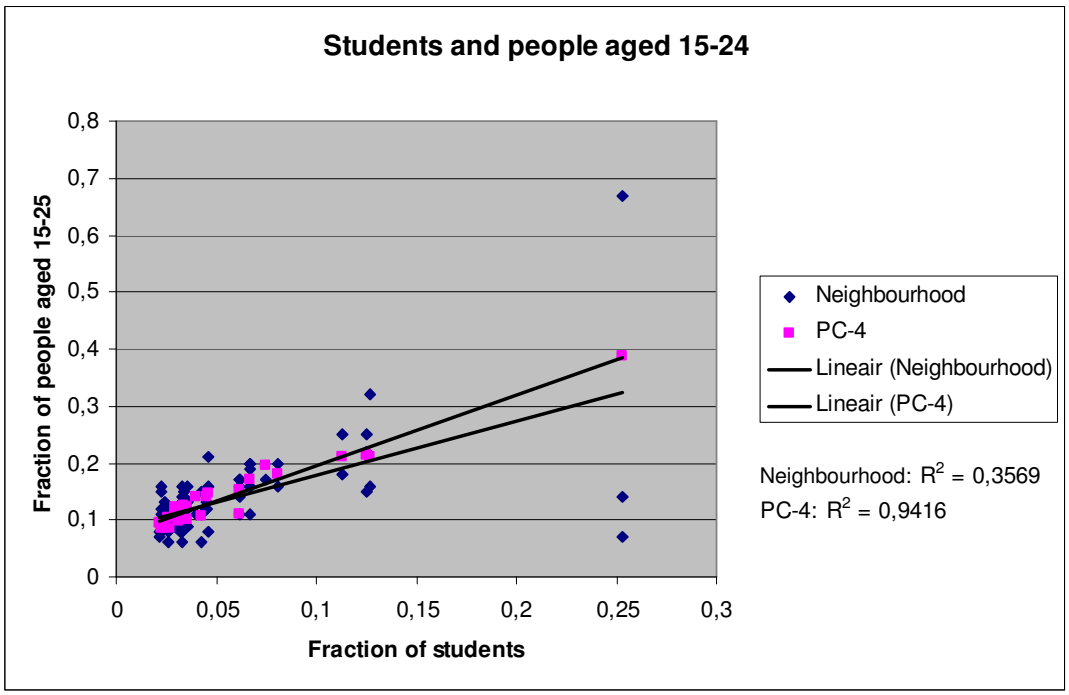


Figure 8.3 The difference between correlation on neighbourhood and PC-4 level

Relative car accessibility of jobs and the number of students
Two other variables that have a somewhat high correlation are the relative car accessibility and the number of students in a zone. Because the correlation is negative a high relative car accessibility (many more jobs accessible by car than by public transport) is accompanied with a low number of students, or a high number of students is accompanied with a low relative car accessibility. Students hence seem to live where the public transport accessibility is better, which means that relative car accessibility is lower there. Because students usually use public transport more than other people this is not unrealistic.

Relative car and bicycle accessibility of jobs
The relative car and bicycle accessibility are also negatively correlated. Zones with higher relative car accessibility hence usually have a worse relative bicycle accessibility. This can be explained because rural areas usually have better relative car accessibility (because less bus and train services are exploited there) while the number of jobs that can be accessed by bicycle are low because the bicycle can only be used to travel small distances, which is an area with low numbers of jobs compared to cities.

Frequency and the number of jobs
The high correlation between the frequency in a zone and the number of jobs (figure 8.4) can be explained by the high frequency in the city centre of Enschede and the corresponding high frequency

in that zone (the upper right point in figure 8.4). When this zone is not taken into account the correlation coefficient becomes 0.349, which is not really high.

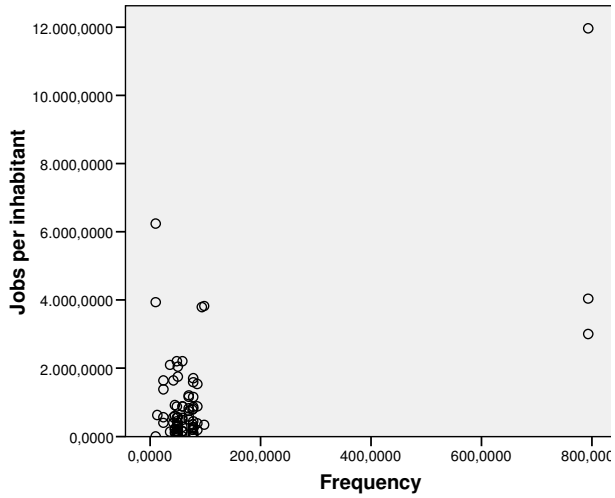


Figure 8.4 The number of jobs per inhabitant plotted against the frequency

Frequency and the number of college enrolments
The correlation between the frequency and the number of college enrolments can also be explained by the high frequency in the centre of Enschede and the large number of college enrolments in this zone. Because college enrolments are only concentrated in a small number of zones including both variables in a model can have positive effects for the model quality.

Punctuality and the relative bicycle accessibility of jobs
The punctuality of services is positively correlated with the relative accessibility of jobs by bicycle. This means that in areas where a highly punctual line is exploited relatively many jobs can be accessed by bicycle in comparison to the number of jobs that can be accessed by public transport. The relative job accessibility by bicycle is high in urban areas because many jobs are located close to each other, which means that many can be accessed by bicycle. Punctuality in the urban area is also higher because more designated infrastructure is built there. This is a possible explanation of the correlation.

Travel time to centre and relative bicycle accessibility
The relative bicycle accessibility is negatively correlated with the travel distance towards the city centre by bus (-0.680). This can be explained because of the decentrality of the neighbourhoods with a long travel time to the centre by bus. Those neighbourhoods also have low numbers of jobs in the vicinity which means that the number of jobs that can be accessed by bicycle is low.

Relative travel time towards the city centre by car and relative car accessibility
The relative travel time towards the city centre by car (compared to the travel time by bus) is positively correlated with the relative car accessibility (coefficient: 0.562). When the relative travel time by car is larger (the travel time by car is relatively long) relatively more jobs can be accessed by car. This is exactly the opposite of what one would expect because the neighbourhoods from which the travel time is relatively long should have more boardings per inhabitant and the neighbourhoods where the relative car accessibility is high should have less boardings per inhabitant. When the correlation with the dependent variable of the two variables is compared it can be seen that the relative travel time by car does not correlate at all with the number of boardings per inhabitants which means that it may be better not to use this variable.

Relative travel time by car and punctuality
Another variable the relative travel time by car towards the city centre is correlated with is the punctuality (-0.554). For trips where the travel time by bus is more competitive to that of the car the punctuality seems to be lower. It is hard to give an explanation for this correlation. One can be that more designated bus infrastructure, which leads to faster trips, does not lead to more punctuality. The number of trips that leave or arrive early can for instance be higher because of this infrastructure. A factor that slightly influences the correlation is that both punctuality and relative travel time are available on a more aggregated scale level than the neighbourhood level, which is used to determine

the correlation. When a more aggregated scale is used the correlation coefficient becomes -0.638, which is even higher. This last factor hence does not lead to a higher correlation. It hence is difficult to explain this correlation, but, because the correlation coefficient is not in the danger zone, no more effort is done to explain the differences.

Relative travel time towards the city centre by bicycle and relative car accessibility
The relative travel time towards the city centre by bicycle (compared to the travel time by bus) is positively correlated with the relative car accessibility (coefficient: 0.597). The relative travel time by bicycle is lowest when busses make detours compared to bicyclists and when bicycling does take very long. It is logical that in areas where this is the case the relative car accessibility is higher because this often is the case in neighbourhoods outside the city centre, where car accessibility is higher.

Relative travel time by bicycle and punctuality
As with the relative travel time by car towards the city the punctuality is related to the relative travel time by bicycle. Again this correlation is hard to understand. Again the scale factor can play a role. Also the fact that reliability of the punctuality numbers has shown to be not perfect (see paragraph 7.6.2) could be an explanation for the correlation.

Relative travel time by bicycle and travel time towards the centre by bus
The relative travel time by bicycle towards the city centre is positively correlated with the travel time towards the centre by bus (coefficient: 0.556). In neighbourhoods where the relative travel time by bicycle is high, the neighbourhoods that are further from the centre, the travel time by bus towards the centre is higher. This correlation hence is as expected.

Address density and relative bicycle accessibility
The address density and the relative bicycle accessibility are positively correlated because the address density and the relative accessibility are both high near the city centre of Enschede and lower further away form the centre.

Address density and distance to stop
The address density is negatively correlated with the distance to stop. The most likely explanation for this negative correlation is the fact that low density neighbourhoods are more spread, and hence larger, than high density neighbourhoods. Because the neighbourhoods are larger the distance to a stop on the line running through the neighbourhood is also larger.

Address density and (relative) travel time towards the city centre
Address density is negatively correlated with the (relative) travel time towards the centre. This can be explained by the fact that the (relative) travel time is higher far from the centre while the address density is higher near the centre.

Population density and the number of inhabitants
The population density and the number of inhabitants in a zone are positively correlated. Zones with high numbers of inhabitants hence also have many inhabitants per square kilometre. This correlation is an effect of the neighbourhood classification by Dutch Statistics. When the 9 neighbourhoods in the centre of Enschede would have been taken together the correlation probably would have been higher, because the population density is highest there and the number of inhabitants is also high there. This randomness in the number of inhabitants suggests that the number of inhabitants is not a valuable variable in a model of the demand for bus transport (as long as the number of boardings per inhabitant is the dependent variable).

Population density and distance to stop
Population is, just as the address density, negatively correlated with the address density. Again the explanation can be found in the fact that more spatial spread leads to higher distances.

Danger zone
As mentioned before the variables that have correlation coefficients that are higher/lower than +/- 0.7 are in the danger zone and, in principle, both variables should not be included in the same regression model.

College enrolments and number of students
The first two variables that have a high positive correlation are the number of college enrolments and the number of students in a zone (coefficient = 0.841). College enrolments are located in only three zones where also many students live. When the correlation of the variables in these three zones alone is determined, it can be seen that there is no clear relation (coefficient = 0.238). The reason for that is that the neighbourhood where most students live is the university and the neighbourhood with most students is the city centre where Saxion and ArtEZ are located. Therefore both variables can be taken into account in the same regression model.

College enrolments and number of jobs
Just as with the number of students, the number of college enrolments has a positive correlation with the number of jobs (coefficient = 0.961). Again when the correlation coefficient of only the three zones with college enrolments are taken into account the correlation coefficient becomes lower (0.552). Because this is outside the danger zone both variables can be included in the same model.

Travel time to the city centre by bus and relative car accessibility
The travel time by bus to the city centre and the relative car accessibility are correlated positively (0.719). This means that the longer the travel time by bus to the centre (the more decentralised a neighbourhood is located) the better the relative car accessibility is. This seems logical because less jobs can be accessed by public transport when a zone is located far from the city centre, because the travel times by PT to the jobs near other buslines are large. Because of this correlation one of the two variables should be chosen as a representative of the accessibility of destinations of the bus.

Relative travel time towards the city centre by car and relative bicycle accessibility
The relative bicycle accessibility is negatively correlated with the relative travel time towards the city centre by car. When the relative travel time by car is high (which means the bus is more competitive) the relative bicycle accessibility should be low. Because the relative travel time by car is higher for neighbourhoods far from the centre of the city the correlation can be explained in a similar way as the correlation with the travel time to the centre by bus.

Relative travel time towards the city centre by bicycle and relative bicycle accessibility
The correlation between the relative travel time towards the city centre and the relative bicycle accessibility is very high (-0.944), which could be expected. The two variables are both measures for the accessibility of destinations of the bicycle compared to the bus. The reason for the negative correlation is that the relative bicycle accessibility is lower for rural areas and neighbourhoods further from the centre where the travel time towards the city centre is higher.

Relative travel time towards the city centre by bicycle and by car
The relative travel times of car and bicycle are highly correlated (0.907). The reason for this probably is that both the relative travel times are higher as the neighbourhood is further away from the centre of Enschede. Taking both variables into account would therefore not lead to a substantially different model with multicollinearity.

Address density and relative car accessibility
Address density is negatively related to relative car accessibility. Just as with the relative bicycle accessibility the reason for the correlation lies in the high address density around the city centre of Enschede. Because the public transport accessibility is high there the relative car accessibility is low and the correlation negative.

Job density and the number of jobs
The positive correlation between the number of jobs and the job density is obvious. When a zone has many jobs the job density, in general, is also high.

Job density and frequency
Both job density and frequency are (extremely) high in the city centre of Enschede. The high positive correlation between the two variables hence is no surprise.

8.2.2 Zones

This time the correlation of the variables in the research objects (zones where the number of boardings is known for) are determined. In appendix G the table with the mutual correlation coefficients is given. The zones around the train station and the University are excluded because the number of boardings in those zones cannot clearly be assigned to the inhabitants in those zones. Also the harbour is excluded because that is a zone where only businesses are located which leads to unrealistic car ownership numbers.
This time the significance of the correlation coefficient is taken into account to determine whether or not the correlation is problematic. For the correlation coefficients it is checked whether or not it is significant to the 0.05 or 0.01 level. The thresholds that were mentioned previously do not perfectly comply with these values. Therefore it is possible that two variables that have a correlation coefficient lower than (+/-)0.5 are considered to correlate with each other.

Interesting findings
The most interesting finding is that many independent variables are correlated with each other on this spatial resolution. In the correlation matrix (appendix G) this can be seen. Most correlating variables have already been discussed based on the correlation coefficients determined on the neighbourhood resolution. Now only the interesting differences with the previously determined coefficients are described.

Car ownership
Car ownership correlates with many other variables, which really contrasts to findings from the earlier correlation calculation. The correlations can however well be explained because the relations with, for instance, business car ownership, income and the number of students have already been proved in literature. The correlation with the variables related to accessibility (relative accessibility of car and bicycle, distance to stop and relative travel time by bicycle) and density can be explained because people living in areas that are well accessible by car and less by bicycle use and therefore own more cars. Also residential self-selection can be present, which means that people are going to live in neighbourhoods that satisfy their wishes. People owning and using many cars are therefore going to live in neighbourhoods where car accessibility is good.

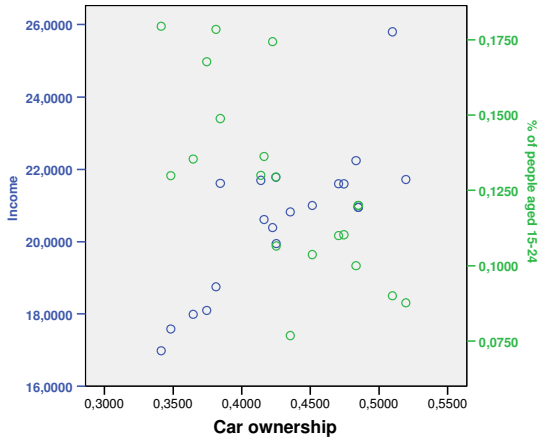


Figure 8.5 Income and percentage of people aged 15-24 as a function of car ownership

Percentage of students/people aged 15-24
The percentage of students is again related to the percentage of people aged 15-24. Because students usually are in that age category this is not a surprise. What was not found before is that especially the number of people aged 15-24 is highly correlated with the (business) car ownership and income. Because people in this age category usually have low wages or do not work full time this correlation is in line with what could expected. The correlation with the variables related to the accessibility of destinations (relative accessibility, relative travel time and distance to stop) can again be explained by the fact that more students live in (urban) areas where the bicycle and bus are relatively attractive in relation to the car. The positive correlation between students and density exists because students usually live in densely built neighbourhoods (with small houses, or many people in larger houses) rather than outskirts that have a less dense built environment.

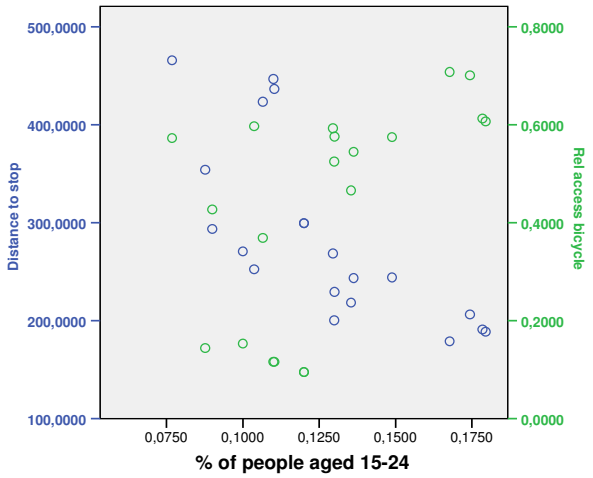


Figure 8.6 Distance to stop and relative bicycle accessibility as a function of the percentage of people aged 15-24

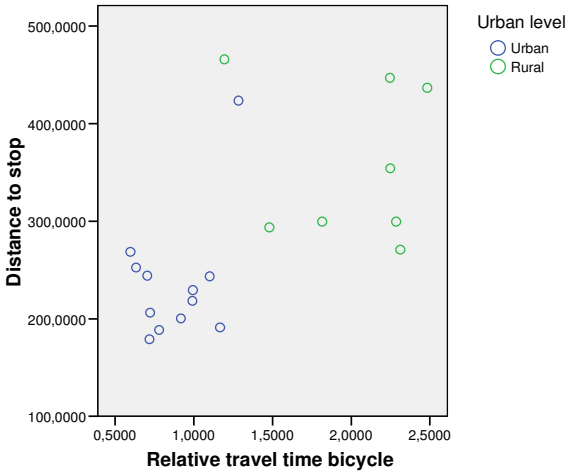


Figure 8.7 Distance to stop as a function of the relative travel time by bicycle subdivided in urban and rural zones

Relative travel time towards the city centre by bicycle
The relative travel time by bicycle towards the city centre is related to nearly all other variables. The travel time by bicycle is relatively long for rural zones. The correlation originates from the fact that some other variables, such as car ownership, number of students, relative car/bicycle accessibility and distance to stop, are also related to the level of urbanization.

Non-correlating variables
The previously mentioned variables all were correlating to many other variables. It is however also interesting what variables are not correlating to other variables, because that are the variables that can be used in the model together. The variables that relate to few other variables are: Income, the number of inhabitants, the number of jobs per inhabitant, frequency, punctuality, travel time by bus towards the city centre and job density.

8.3 Conclusion

The previously executed calculation of the correlation coefficient and the explanation that was given can now be used to make choices on the variables to exclude from the research. Groups of variables that correlate with each other are described together.

8.3.1 Accessibility of destinations

Different variables were included to represent the accessibility of destinations by bus. With the regional model the relative accessibility was determined based on the number of jobs that could be accessed by car, bicycle and public transport as a whole. Because public transport as a whole is used and train could become the dominant mode it was chosen to also look at the relative travel time towards the centre of Enschede with the different modes (car, bicycle and bus). Because the mutual correlations (see figure 8.8 for an x-y plot) are high between these variables they should not all be taken into account in the regression at the same time.

When the correlation of the different variables with the dependent variable (table 8.2) is taken into account it can be seen that

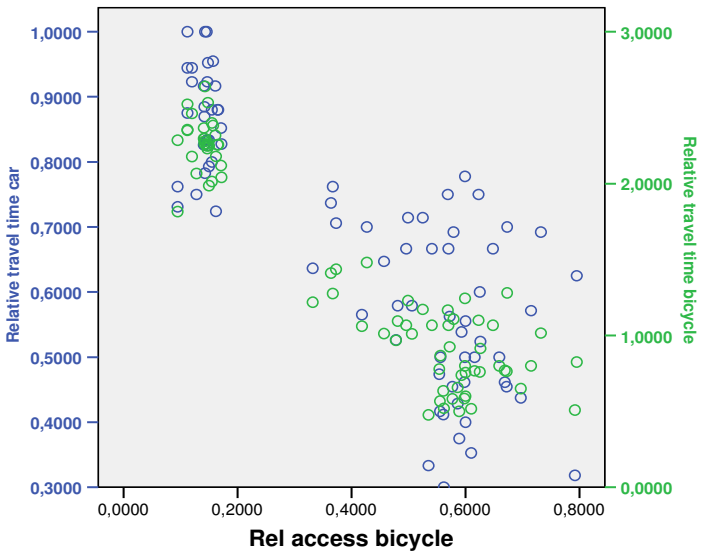


Figure 8.8 Relative travel time by car and bicycle as a function of the relative bicycle accessibility

relative car accessibility has a low correlation coefficient. This indicates that no real relation could be found between the independent and the dependent variable and that it is better not to start the regression with this variable. Because the relative bicycle accessibility is determined based on the same model this should also not be the initial measure for the accessibility of destinations by bus. The other variables (relative travel time towards the city centre) therefore is used first during the regression. The first is travel time towards the city centre by bus because this variable correlated with few other variables.

8.3.2 Spatial density

Looking at the spatial density (population or address density), it can be seen that the correlation between the address density with the number of boardings per inhabitant is highest. The two variables have the same number of variables they are correlated with. The address density therefore seems the most promising variable, but also the population density is taken into account during the regression, however not simultaneously with the address density.

8.3.3 Jobs

As expected the number of jobs per inhabitant and the job density are highly correlated. Both variables are not correlated to many other variables (three). The job density seems to be somewhat more correlated to the number of boardings per inhabitant and is therefore used first in the regression.

8.3.4 Number of students

Both the relative number of students and the percentage of people aged 15-24 stay included in the regression to find out what implications it has to use more detailed, but not directly measured, measures of the number of students in a zone/neighbourhood. Both variables are however not be used simultaneously.

8.3.5 Number of inhabitants

The number of inhabitants in a zone is already taken into account in the dependent variable, because the number of boardings per inhabitant is used. It is however possible that the number of inhabitants in a zone does influence the number of boardings per inhabitant. It is however doubtful whether or not the number of inhabitants in a zone can be used as a determinant of the demand for bus transport. The zones that are the research object in this research are dependent on the lines that run through them and the number of stops where counts are undertaken. The size of the zone, and related to that the number of inhabitants, hence is very random and arbitrary. When a model is used with the number of inhabitants as a determinant for the number of boardings per inhabitant in a zone the number of boardings can be influenced by making the zones larger or smaller. Dependent on the sign of the parameter increasing zonal size either increases or decreases the number of boardings in total. When the parameter for instance is -0.0001 and two zones of 200 inhabitants are used the influence on the number of boardings per inhabitant is -0.02 for each zone while taking the zones together leads to an influence of -0.04. The influence on the total number of boardings when the zones are considered to be two zones then is -8 and when the zones are taken together it is -16 which. Because of this the number of inhabitants is not included in the regression models.

9 Regression

In the correlation chapter conclusions were drawn about which variables could be good determinants for the number of boardings per inhabitant and which variables should and should not be used in a model together. In this chapter the next step in the research, the regression, is described.

The goal of the regression is to find a (small) group of variables that explains a large part of the variance in the number of boardings per inhabitant in the different zones.
Based on the correlation coefficients between the different variables first some possible determinants are indicated that are used in the first regression step. After that variables that are not significant are excluded or comparable or extra variables are included to check whether the explained variance or the significance level of the parameters increases.
After that it is discussed what model is the best representation of the reality. This can differ from the variables in the model that explains the variance of the number of boardings per inhabitant best.
In the last part of the chapter models are estimated in which the university and harbour area are included to estimate the models.

9.1 Design

To keep the model as simple as possible linear regression is used. This means that the dependent variable is linearly related to the different variables. The model then looks like:

$$Y = X_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \qquad (9.1)$$

Where:
Y = The number of boardings per inhabitant
X_0 = A constant that represents the number of boardings when all variables are 0
β_i = The parameter value of variable i
X_i = The value of variable i

As with the correlation not all zones are included in the regression. Three zones are excluded: The city centre of Enschede, the university and the harbour area in the West of Enschede. The city centre is excluded because the boardings in that zone cannot be attributed to the people that live there. The university is excluded because it is the only other zone where a substantial amount of college enrolments are located and people boarding there in many cases are not the people that live in the zone. The harbour area is excluded because nearly nobody lives there and many cars are owned by the companies. All these zones would reduce the quality of the model because the values of the different variables are extreme in relation to the values of the other zones.

9.2 Initial regression

With the outcomes of the correlation analysis the first regression step is undertaken. The variables included and the outcomes are described below.

9.2.1 Variables included

From the correlation with the number of boardings per inhabitant it can be concluded that the frequency of services is an important determinant that should always be included in the model.
Because students are important bus users one of the two variables concerning the number of students in a zone also has to be included. Because the percentage of people aged 15-24 is most detailed, has the right sign of the correlation coefficient and has higher correlation than the number of students this variable is included first.
Also the income per household and car ownership is included in the first model. This is done because both variables show a relatively high correlation with the dependent variable and because both variables were used often in other research.
The accessibility of destinations is represented by the travel time by bus towards the centre because the correlation is highest. Also the punctuality of the line is included because this variable correlates little with other variables.

9.2.2 Outcomes

In table 9.1 an overview of the outcomes of the regression is given. The (adjusted) R^2 of the model is 0,742 which means that about 74% of the variance is explained by the model. The model should be as close as possible to a R^2 value of 1.00, which would mean that the model explains all variance and has a perfect fit to the data.

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta	B	Std. Error
(Constant)	-0,234	0,093		-2,512	0,026
% of people aged 15-24	0,154	0,175	0,175	0,879	0,395
Car ownership	0,330	0,161	0,656	2,052	0,061
Frequency	0,00126	0,00021	1,080	6,084	0,000
Income	-0,00183	0,00288	-0,137	-0,637	0,535
Punctuality	0,00128	0,00061	0,285	2,098	0,056
Travel time by bus	0,00062	0,00093	0,112	0,664	0,518

Table 9.1 Regression outcomes initial model

In table 9.1 the (un)standardized coefficients, the standard error, the T-value and the significance level are presented. The unstandardized coefficient (β in table 9.1) of the frequency for instance is 0.00126, which means that an increase of the frequency per day by 1 leads to an increase of the number of boardings per inhabitant of 0.00126. When the increase of the frequency is 20, the increase in the number of boardings is 0.02320 (20 * 0.00126). The standard error is 0.00021 which is very low compared to the coefficient. The standardized coefficient gives an indication of the relative importance of the different coefficients. The value of 1.08 for the frequency indicates that it is about 4 times as important as punctuality (0.28). The t-test is used to determine the significance level of the value of the coefficient, with higher values of t meaning a higher significance level. A t-value[14] (20 cases. 19 degrees of freedom) of 2.093 is associated to a significance level of 5% and a value of 2.861 with a significance level of 1%. A significance level of 5% means that it is for 95% sure that the value of the parameter is significant. When the text refers to a higher significance level, this means that the certainty about the significance of a parameter is larger. A significance level of 5% hence is higher that a significance level of 10%. The value of 6.08 for the frequency corresponds to a significance level of 0.00004.

It can be seen that three of the variables are clearly not significant. This are: Income (0.535). Travel time by bus towards the centre of Enschede (0.518) and Percentage of people aged 15-24 years (0.395).
The low significance level of the income can be explained by the high correlation between the income and car ownership. Richer people tend to have more cars and using both variables therefore can lead to an insignificant relation with one of the variables.
The travel time by bus was one of the measures for the relative attractiveness of the bus. Because this variable does not have a significant relation in this model, another measure can be used. The same accounts for the number of people aged 15-24 years that can be replaced by the (real) number of students.

9.3 Next steps

With the findings from the first model more models are computed. In appendix H an overview of the model parameters of these models is given. In this paragraph the models themselves and their consequences for the explained variance and the significance level are described.

---

[14] $t = \frac{\hat{\beta}}{SE_{\hat{\beta}}}$

With: $\hat{\beta}$ = the coefficient estimate

$SE_{\hat{\beta}}$ = the standard error of the estimate

Model 2: Without income, TT by bus replaced by relative TT by car
Because the income and the travel time by bus have the lowest significance level, these two variables are excluded from the regression analysis. To include a measure of the relative attractiveness of the bus the relative travel time by car is included in the next step of the regression. The effect is that the R^2 increases to 0.750. Still however two variables, the percentage of people aged 15-24 and the relative travel time by car towards the city centre of Enschede, have a quite insignificant parameter value. Model 2 however now is the model with the highest explained variance.

Model 3: % of people aged 15-24 replaced by % of students
Because the percentage of people aged 15-24 had an insignificant coefficient earlier this variable is replaced by the percentage of students. Doing so leads to an explained variance of 0.735, which is lower than before. The insignificance of the parameter of the percentage of students is also higher than the insignificance of the percentage of students earlier (0.66 to 0.32). Also the insignificance of the relative travel time by car is higher than in the previous step. Switching from the percentage of people aged 15-24 to the percentage of students hence reduces the fit and significance of the model.

Model 4: Model 2 with relative TT by car replaced by relative TT by bicycle
In the fourth model the relative travel time by bicycle instead of the relative travel time by car is used and, as in model 2, students are represented by the number people aged 15-24. This model also explains the variance less than model 2 and the parameters have lower significance levels. The relative travel time by bicycle hence is a worse determinant than the relative travel time by car.

Model 5: Excluding the accessibility of destinations
Another option is to only use punctuality as a measure for the attractiveness of the bus and not use a measure for the accessibility of destinations. The (relative) travel time towards the city centre therefore is excluded in model 5. Doing so leads to an increase of the explained variance to 0.755. Model 5 hence now explains the variance of the dataset most.

Model 6: Car ownership replaced by business car ownership
Although car ownership is a significant variable in the model it is possible that including business car ownership leads to a better fit. In model 6 therefore car ownership is replaced by business car ownership. This decreases the $R^2$of the model to 0.657 and the business car ownership parameter has a low significance level. Another negative consequence is that the insignificance of the coefficient for percentage of people aged 15-24 increases. Business car ownership hence is a worse determinant than “normal” car ownership.

Model 7: Model 5 without the percentage of people aged 15-24
Although the percentage of students is thought to be an important determinant of the demand for bus transport, the significance level of the parameter value is not high at all in the previous models (32-52%). The variable is therefore excluded in model 7. This leads to a slightly higher explained variance (0.757) with variables that all have a significance level higher than 3%.

Model 8: Including distance to stop
In model 8, model 7 is expanded with another variable that can be used to represent the attractiveness of the bus, the average distance to a stop. Doing so reduces the explained variance (R^2 = 0.744). The coefficient for the distance to a stop is also not significant at all (significance level is 69%). The distance towards a stop hence is not a good determinant of the demand for bus transport, with the variables in this model.

Model 9: Model 7 with relative car accessibility
Earlier it was concluded that the three variables about the (relative) travel time towards the centre were no good determinants of the number of boardings per inhabitant. Two other variables related to the accessibility of destinations by bus are the relative accessibility by car and bicycle. A lack of clear correlation between these variables and the dependent variable and the uncertainty about the real applicability of this variable was a reason to not use these variables in first instance. It is however possible that the variables substantially improve the model. To check this, the relative accessibility by car is included in this model which leads to a substantial increase of the explanatory value of the model (R^2 = 0.784). A disadvantage is that the significance level of car ownership decreases (18%) and that the significance level of the relative accessibility by car could be higher (10%). Another problem with the measure is the positive sign of the parameter. Increased car accessibility hence

leads to increased bus use, which is not according to the expectations. Although this model explains the variance in the dataset best so far, the variable is not used in the upcoming models. Reason for this is the large influence of train availability on this variable, the low significance levels and the unexpected parameter sign.

Model 10: Relative car accessibility replaced by relative bicycle accessibility
The last possible measure of the accessibility of destinations is the relative bicycle accessibility. This variable replaces the relative accessibility by car in model 10. The results are however disappointing, because the explained variance reduces (R^2 = 0.742) and the significance level of the new variable is extremely low (85%). The relative accessibility by bicycle is also not useful in a model for the number of boardings per inhabitant.

Model 11: Model 7 with job density
Until now no measure for the presence of jobs in a zone was used in a model. In model 11 the job density, which has most correlation with the number of boardings per inhabitant, is included. Disappointingly the explained variance reduces by including this variable (R^2 = 0.741). The significance level of the new variable is also very low (97%). Including this variable hence does not lead to a better model.

Model 12: Job density replaced by jobs per inhabitant
Since the job density does not lead to a better model another measure for the presence of jobs in a zone is used, the number of jobs per inhabitant. The explained variance of this model is slightly higher, but still lower than model 7. The significance level of the variable is also higher than the significance level of the job density (71%), but still way too low. The number of jobs per inhabitant hence is a better determinant than the job density, but with these variables included the model is still not better than model 7.

Model 13: Without car ownership, with address density
In the previous models the sign of the coefficient for car ownership is positive. Literature however proves that car ownership has a negative relation to the demand for bus transport. The more cars are owned in a zone the less people should make use of the bus. The variable therefore is excluded from this model. Together with that a measure for the housing density is included. Because the address density showed the largest correlation with the number of boardings per inhabitant the address density is used as a variable in this model. The result is that the R^2 increases to 0.798 with pretty high significance levels of all variables (18% and higher, two-tailed). This model hence has a higher explained variance than the one that is currently highest (7), and also has variables with a one-tailed significance level higher than 10%.

Model 14: Address density replaced by population density
The explained variance of model 13 is quite high. The question however is whether replacing the address density by the other measure of housing density, population density, is beneficial for the model. Therefore population density is included as a variable in model 14 replacing address density. The explained variance of this model is a lot lower than the model with address density (R^2 = 0.685). The significance levels of the different parameter values are lower than for model 13 with the 61% of the number of jobs per inhabitant as an extreme example. The population density hence is a worse determinant of the number of boardings than the address density.

Model 15: Model 13 with % of people aged 15-24
Because the number of students in a zone stays important for bus the percentage of people aged 15-24, which was found to be a better determinant of the number of boardings than the number of students earlier, is reintroduced in the model. Interestingly the explanatory value of the model this time increases (R^2 = 0.804). As previously the variables' parameter is not extremely significant (24.6%), but it is better than in the earlier models. This previous low significance levels possibly are the result of the correlation between car ownership and the percentage of people aged 15-24. The explained variance of the model is the highest so far, but not all variables are as significant as desired.

Model 16: Without jobs per inhabitant
The number of jobs per inhabitant has, apart from the percentage of people aged 15-24, the lowest significance level in model 15. Therefore this variable is left out of model 16. Doing so results in a lower explained variance (R^2 = 0.779) and a lower significance level of the parameter value of the

percentage of people aged 15-24. Excluding the number of jobs per inhabitant hence leads to a worse model.

Model 17: Model 13, with relative travel time by car
The last models did not include a measure of the accessibility of destinations. This variable however still is thought to be an important determinant of the demand for bus transport. The measure with the highest significance level of the different variables related to the accessibility of destinations was the relative travel time by car, which is therefore included in model 17. The model has a slightly higher explained variance than model 13 and a slightly lower explained variance than model 15 (R^2 = 0.799). The significance level of the different variables however decreases and the significance level of the relative travel time by car is also not high (32.3%).

Model 18: Without jobs per inhabitant
The variable that has the highest standard error in the significance in model 17, apart from the relative travel time by car, is the number of jobs per inhabitant. In model 18 this variable is therefore is excluded. The exclusion of this variable leads to a slightly lower explained variance of the model (R^2 = 0.787). The significance level of the different variables also decreases. This model hence is worse than model 17.

9.4 Best model(s)

From the 18 different models described above a model should be chosen that is used as the model of the demand for bus travel. The definitive model should explain a large part of the variance of the data and should have parameter values that have a high significance level. Also no variables that could lead to a bias because of the scale level or background of the variable should be included. The sign of the parameter should also be explainable. Besides these criteria related to the appropriateness of the model another important criterion is the ease of obtaining the data, because obtaining the data should not take too much effort.

Significance level
Significant parameter values of variables are very important for a valuable model. Ideally all parameter values can be said to be significant with 100% certainty (significance level of 0%). Having such tight requirements would however lead to very few variables that can be used that explain very little of the variance in the dataset. Looser, but still tight, requirements are (one tailed) significance levels of 5% and 10%. When the 5% requirement is used (only significance levels marked green and yellow in appendix H) only one model (7) can be used and when this is loosened to 10% (also including the orange marked values) two more models (9 and 13) can be used. Although the significance level of all parameters is not very high, model 15 is included as a possible model. The reason for that is that this is the model where the number of students has the highest significance level (12.3%, one-tailed). The significance of the parameters is however negatively influenced by the low number of research objects (cases). When more research objects are present the parameter values usually have a higher significance level. Because the number/percentage of students is thought to be an important parameter this model is included in the comparison of the different models.
The different models and the variables that are included are displayed in table 9.2. In this table the cells that are marked green indicate a one-tailed significance level of 1%, yellow indicates a 5% significance level, orange indicates a 10% significance level and red indicates a significance level higher than 10%.

Model	Adjusted R^2		Constant	% of people aged 15-24	Car ownership	Frequency	Jobs per inhabitant	Punctuality	Rel access car	Address density
7	0.757	Coefficient	-0.20924		0.24270	0.00124		0.00137		
		Std. Error Sign.	0.005		0.008	0.000		0.025		
9	0.784	Coefficient	-0.20333		0.13759	0.00118		0.00158	0.00637	
		Std. Error Sign.	0.004		0.176	0.000		0.010	0.104	
13	0.798	Coefficient	-0.10171			0.00101	0.01792	0.00183		-0.00002
		Std. Error Sign.	0.010			0.000	0.185	0.004		0.002
15	0.804	Coefficient	-0.11670	0.00161		0.00099	0.02293	0.00185		-0.00002
		Std. Error Sign.	0.006	0.246		0.000	0.108	0.004		0.003

Table 9.2 potential models and their coefficient values

In figure 9.1 the model values for the number of boardings per inhabitant are displayed. It can be seen that the spread of the models is different. A larger spread leads to a lower the value of R^2, so model 15 has the smallest and model 7 the largest spread. Another interesting finding is that the different models give comparable outcomes (the values are close to each other). In most cases all models either over- or underestimate the number of boardings.

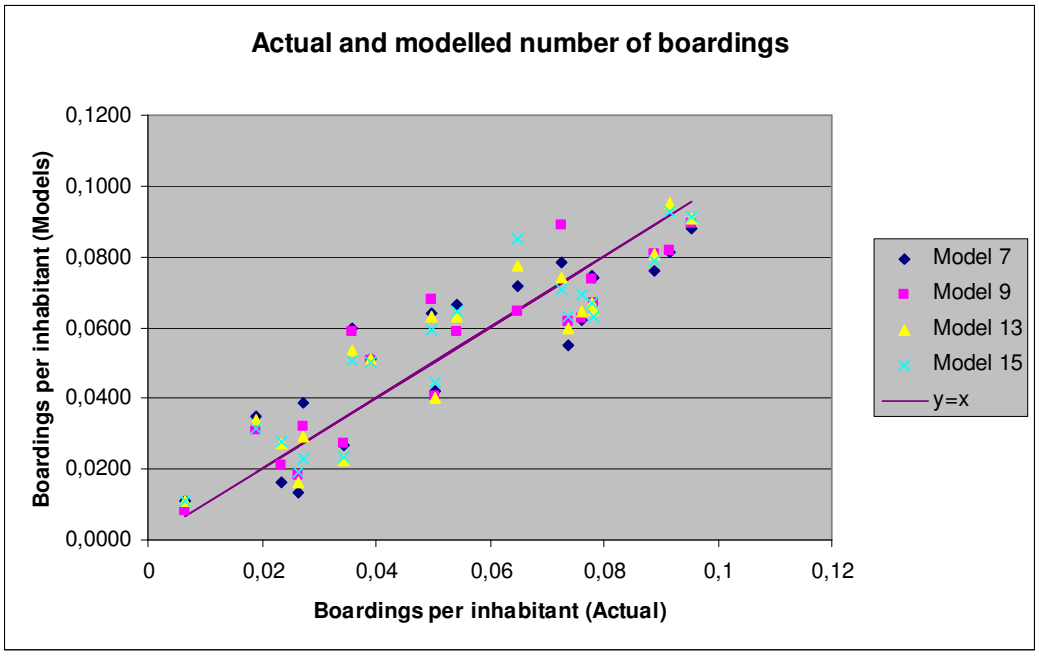


Figure 9.1 Comparison of the actual number of boardings with the model outcomes

Explanation of variance

The different models have different values of R^2, which is the measure for the variance that is explained by the model. Model 15 explains the variance best followed by model 13, 9 and 7. The explained variance of model 7 is however still high (R^2 = 0.757).

Variables background

In the description of model 9 (in the previous paragraph) it was already stated that the relative car accessibility of jobs could not be used as a determinant. The problem lies partially in the background of the variable. The relative car accessibility is computed by dividing the number of jobs that can be accessed within reasonable time by car by the number of jobs that can be accessed by public transport within reasonable time. Public transport however is not the same as just bus transport, because train is also very important for the public transport accessibility.

Parameter sign
The signs of the parameters of the different variables are the same for all models included in the choice process. This means that no strange differences between the models are present.
The sign of the constant is negative. A certain minimum value of the other variables hence is needed to get a positive number of boardings per inhabitant. This can be explained by the fact that at least a certain frequency and punctuality should be present for people in a certain zone to use a certain line. When frequency and punctuality are low one can expect that people are not using the bus at all because of the inconvenience of travelling with it. Because of the negative sign of the constant, the total number of boardings stays (below) zero which corresponds to the expectations. The negative sign of the constant hence is as could be expected. When the constant would have had a positive sign the number of boardings is positive even when there are no services, which of course is impossible.
The sign for the percentage of people aged 15-24 (model 15) is positive. This also corresponds to expectations because larger fractions of students should lead to higher numbers of boardings.
Car ownership (model 7, 9) has a positive sign. This should mean that the more cars are owned per inhabitant the more trips are undertaken by bus. This is strange because both modes (car and bus) are substitutes and people that own more cars on average use the bus less (van Acker & Witlox, 2010, Paulley et al, 2006 & Maat & Timmermans, 2009). An explanation can be that the overall mobility (number of trips) is higher for people with higher car ownership, for instance because car ownership is also correlated to the employment status of people. People that have a job have higher mobility needs, because they have to travel to their work. When some of the extra trips by these people are be made by bus their bus use is higher than people without a job and a car. Another explanation can be that car ownership per inhabitant is higher in industrial/business areas, where the number of boardings per inhabitant is also high (because few people live there and other people board in those zones on their way home). If the last explanation is true the use of car ownership as a variable in the model would not explain the differences in bus use that are accompanied by car ownership, but the differences accompanied by the differences in land use (business or residential).
The sign of the parameter for the frequency and punctuality are both positive which corresponds to the expectations. When the frequency and punctuality are higher the number of boardings per inhabitant is also higher.
The parameter value for the number of jobs per inhabitant is positive. When more jobs per inhabitant are located in a zone the number of boardings is higher. This is in line with the expectations because a larger number of jobs also leads to an increase in boardings of people from work to home.
The relative car accessibility parameter has a positive sign. When relatively more jobs can be accessed by car more trips are taken by bus. This is not in line with the initial expectations, which stated that the bus is relatively more interesting when relatively more jobs can be accessed by bus than by car. Because the sign is different than expected, it is not allowed to use the one tailed significance level. The two-tailed significance level of the relative car accessibility is 10.4%.
The parameter value for the address density is negative. As from Balcombe et al. (2004) one would however expect that the number of boardings per inhabitant is higher in more densely built zones. The fact however is that the number of boardings per inhabitant is higher in the areas further away from the centre (see figure 7.8) therefore this sign is explainable.

Ease to obtain data
Most data in the models are easy to obtain. People aged 15-24 and car ownership can be found in the neighbourhood map, the frequency comes from the schedule, the punctuality is reported quarterly and the number of jobs and inhabitants are included in the regional model. The only variable that is hard to obtain is the relative accessibility, which is computed using the regional model.

Choice
From the five models one should be chosen. Model 9 is the first one that is excluded because the relative car accessibility is included in this model. The relative car accessibility has problems with the variable background, has an unexpected sign (leading to a low significance level) and is hard to obtain. Also model 7 is not chosen because the car ownership is included in the model, which has an unexpected sign that could not be explained. The other models (13 and 15) have increasing explanatory value but decreasing significance of the parameter values. When a one-tailed 10% error margin is strictly used model 13 would be the only possible model. One variable in model 15 has a significance level of 12.6%, the advantage of this model however is that the number of people aged 15-24 (a measure for the number of students) is included and that the explained variance is higher. The low significance level also can be related to the low number of cases included in the regression. The model therefore is not excluded before the validation process. In the validation chapter (10) the

two models are tested with data from another year to determine which model works best for both years.

Implications of the coefficient values
To give an indication of the effects of the different variables on the number of boardings the different parameter values of the variables in model 15 (table 9.3) is described briefly.

Model 15 (R^2 = 0.804)	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta	B	Std. Error
Constant	-0.117	0.036		-3.230	0.006
% of people aged 15-24	0.00161	0.00133	0.183	1.212	0.246
Address density	-2.46E-05	6.70E-06	-0.593	-3.669	0.003
Frequency	0.00099	0.00013	0.852	7.869	0.000
Jobs per inhabitant	0.0229	0.0134	0.188	1.716	0.108
Punctuality	0.00185	0.00054	0.411	3.441	0.004

Table 9.3 Model parameters of model 15

Constant
The value of the constant is -0.117. This means that the other variables have to have some base value in order to get a positive number of boardings (per inhabitant).

Percentage of people aged 15-24
The parameter value for the percentage of people has a value of 0.00161. When the percentage of people aged 15-24 in a zone increases from 10% to 20%, the number of boardings per inhabitant increases with 0.0161. For a zone with 1000 inhabitants this means an increase of 16.1 boardings a weekday. The range of the values is between 8.8% and 17.9%. The maximum difference in boardings per inhabitant then is 0.0147.

Address density
The parameter value for the address density is -0.0000246. This value is much lower than the one for the percentage of people aged 15-24 because the value of the variables is very different. Address density for instance ranges from 283 to 2439 addresses within a radius of 1 kilometre, while the proportion of people aged 15-24 ranges from 0.088 to 0.179. When the address density is increased with 1000, the number of boardings per inhabitant decreases with 0.0246, which is 24.6 boardings in a zone with 1000 inhabitants. The maximum difference in the number of boardings per inhabitant is -0.0530.

Frequency
The frequency has a parameter value of 0.00099. This means that an increase of the frequency by 24 services a day (one per hour) leads to an increase of the number of boardings per inhabitant of 0.0238 and an increase of the number boardings of 23.8 in a zone of 1000 inhabitants. Because the range of the frequency is 13 to 78 services a day, the largest difference in the number of boardings per inhabitant is 0.0644.

Jobs per inhabitant
The parameter of the number of jobs per inhabitant has a value of 0.0229. This variable has values in the range of 0.056 to 0.830. The difference between the number of boardings per inhabitant in the zone with the highest and the lowest value therefore is 0.0177.

Punctuality
The parameter value of the punctuality is 0.00185. Because the range of the punctuality of the different lines is not very wide (54.988-78.150), the maximum difference in the number of boardings per inhabitant are also not very high (0.0428). This maximum difference is however larger than the maximum difference following from the number of jobs per inhabitant, which indicates that punctuality is relatively more important.

Relative importance
The relative importance of the variables is shown in the fourth column. When a value of Beta is twice as high as another Beta the variable is twice as important. In table 9.3 it can be seen that the frequency is most important. The importance of the address density is about 70% of the importance of the frequency and the other variables are 20% (students and jobs) and 50% (punctuality) as important as the frequency.

9.5 Including the university and the harbour area

Ideally all zones are included in the model estimation because the model then also can be used for all zones. In the previous models three zones, the city centre, the university and the harbour, were however not taken into account. Extreme values of the number of boardings per inhabitant and/or car ownership were the reasons for that. The values in the city centre are really too extreme, but the other zones can probably be included in the regression phase to compute a model that also is applicable to business zones and zones with many college enrolments.

University only
First a regression is done where only the university zone is added to the zones used before. Because college enrolments are very important in this zone the enrolments are included as one of the variables. The other variables are the same as in the definitive model that was determined in the previous paragraph. The models can be seen in table 9.4.

Model	Adjusted R^2		Constant	% of people aged 15-24	Frequency	Jobs per inhabitant	Punctuality	Address density	College enrolments per inhabitant
u13	0.955	Coefficient	-0.10123		0.00101	0.01808	0.00182	-0.00002	0.05547
		Significance	0.011		0.000	0.185	0.005	0.002	0.000
u15	0.957	Coefficient	-0.11718	0.00168	0.00100	0.02326	0.00184	-0.00002	0.02520
		Significance	0.006	0.225	0.000	0.104	0.004	0.002	0.322

Table 9.4 Model parameters with the university included

Both models explain a(n extremely) large part of the variance of the data. Most likely a large part of the extra fit compared to the models without the university is the large deviation of the number boardings per inhabitant between the university (0.288) and the other zones (range: 0.007-0.096). The university therefore is very important and the model has low deviation in the university zone and high deviation in the other zones that are relatively less important. The significance levels of the different models are comparable to the significance levels for the models without the university as a zone. In model u15 it can be seen that the percentage of people aged 15-24 and the number of college enrolments are correlated, mainly because many students live on the campus of the university. The significance level of the percentage of people aged 15-24 however is higher (22.5% compared to 24.6%) that in model 14.
Interesting is that the parameter values of the different variables are nearly the same for the different models. This indicates that the “normal” model is still valid and the number of college enrolments is used to explain the differences between the university zone and the other zones.
In figure 9.2 it can be seen that the number of boardings per inhabitant in this zone is much higher than the others, which does influence the value of R^2 hugely. Because the parameter values are nearly the same as for the models without the university zone the explained variance for the other variables are about the same as without the university and the number of college enrolments.

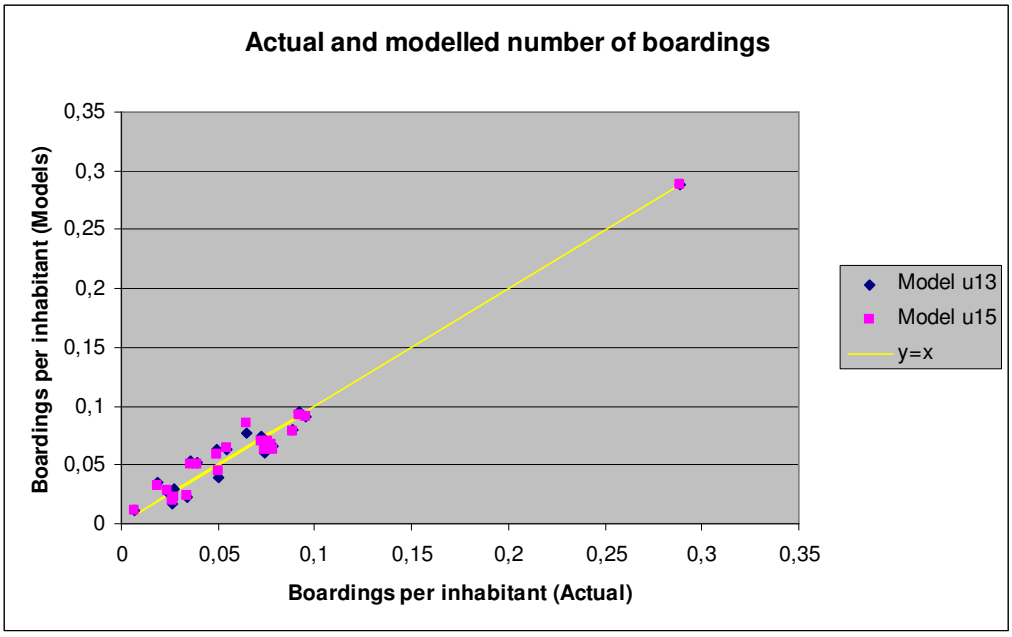


Figure 9.2 Actual number of boardings and model outcomes with the university zone included

University adjusted

To overcome the problems with the extreme value for the number of boardings per inhabitant two extra models are estimated. For these models it is assumed that all students make the same number of trips as all people living in the university zone. Because the main interest is the number of trips of people living in the area the total number of boardings is multiplied by the fraction of people living on the university zone over the total number of students, see formula 9.1.

$$B_{UI} = \frac{B_U}{I_U} * \frac{I_U}{S} = 0.2884 * \frac{2615}{8587} = 0.0878 \qquad (9.1)$$

Where:

B_{UI} = The number of boardings per university zone inhabitant

B_U = The number of boardings at the university zone

I_U = The number of inhabitants of the university zone

S = The number of students at the university

The number of college enrolments per inhabitant now is excluded as a dependent variable because the change in of the number of boardings already is used to correct for the number of college enrolments. In table 9.5 the model parameters of the two models are presented.

Model	R^2		(Constant)	% of people aged 15-24	Frequency	Jobs per inhabitant	Punctuality	Address density
ua13	0.805	Coefficient	-0.09347		0.00100	0.00903	0.00173	-0.00002
		Significance	0.012		0.000	0.303	0.005	0.002
ua15	0.797	Coefficient	-0.09693	-0.00020	0.00100	0.01412	0.00179	-0.00002
		Significance	0.013	0.560	0.000	0.264	0.005	0.003

Table 9.5 Model parameters with the adjusted university zone included

The explained variance for these models is back to “normal” values. Disappointingly the significance levels of the different parameter values are lower than for the models with and without the university zone. Interesting is the fact that the parameter value of the percentage of people aged 15-24 is

negative in model ua15. A higher number of students in a zone hence leads to a lower number of boardings per inhabitant, which is really unexpected. The explained variance of model ua15 also is lower than for the model without students included and its parameters have a lower significance level, which indicates that the inclusion of the variable is not beneficial for the model.
In figure 9.3 the actual and modelled number of boardings per inhabitant are visualised. It can be seen that the estimates of the models, ua13 and ua15, in most cases are very close to each other. This means that including the percentage of students does not really alter the estimation. Another interesting outcome is that the model estimates do not seem to differ that much from the estimates done with the models u13 and u15, so without the adjustments made. No real improvements hence are made by adjusting the university zone.

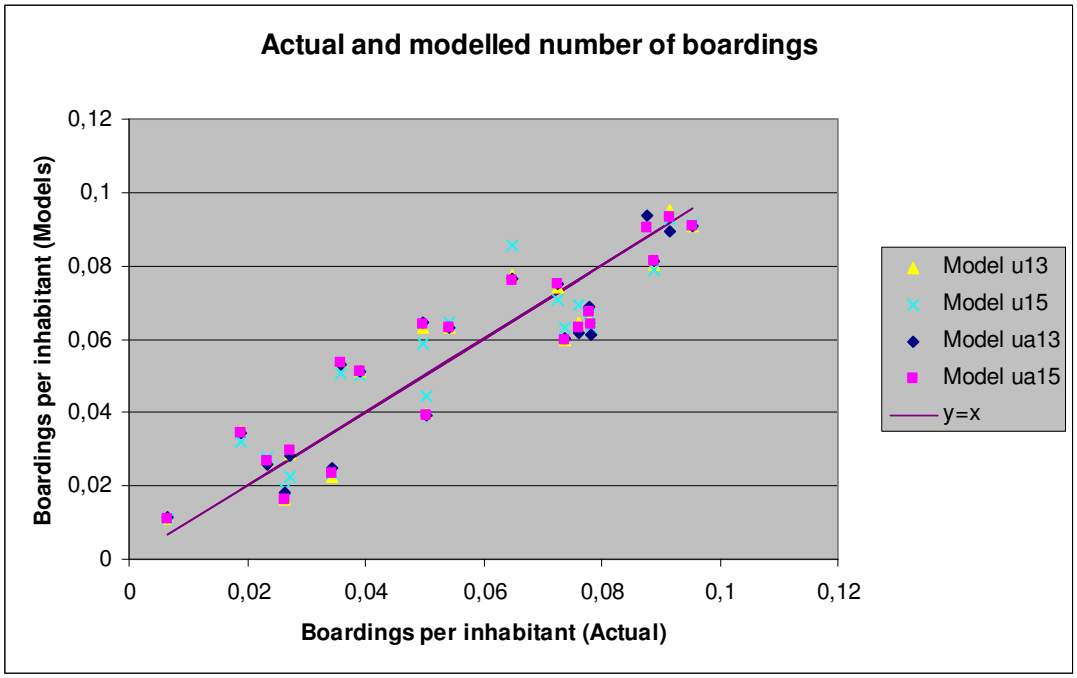


Figure 9.3 Actual number of boardings and model outcomes with the adjusted university zone

Harbour only
In this step only the harbour zone is included in the regression. This time college enrolments are not taken into account because no education institutes are located in the harbour zone. In table 9.6 the model parameters are displayed.

Model	R^2		Constant	% of people aged 15-24	Frequency	Jobs per inhabitant	Punctuality	Address density
h13	0.82	Coefficient	-0.09066		0.00103	0.00286	0.00170	-0.00002
		Significance	0.016		0.000	0.000	0.006	0.002
h15	0.815	Coefficient	-0.09765	0.00099	0.00102	0.00287	0.00169	-0.00002
		Significance	0.014	0.462	0.000	0.000	0.008	0.005

Table 9.6 Model parameters with the harbour included

The explained variance is somewhat higher than the “normal” models 13 and 15. The significance level of the percentage of people aged 15-24 is however a lot lower in model 15 than in model h15. An important notice to make is that the value of the number of jobs per inhabitant is extremely high in the harbour zone (32.519) compared to the other zones (range: 0.056-0.830). Because of this extreme value it is obvious that the significance level of the parameter for the number of jobs, which was quite low in models 13 and 15, is very high. When only the explained variance and the significance levels are taken into account it can be said that model h13 is better than model 13 and that model 15 cannot

be used because of the low significance level of the parameter for the percentage of people aged 15-24.
In figure 9.4 the model outcomes and the actual number of boardings per inhabitant are shown. It can be seen that the models give a reasonably good estimate of the number of boardings per inhabitant. The harbour zone is the zone with most boardings per inhabitant. Model h13 seems to be really calibrated to reproduce this value because the estimate is really close to the actual value. Including the percentage of people aged 15-24 increases the estimate, and consequently the deviation from the real value. For the other zones the estimates of the two models do not differ a lot.

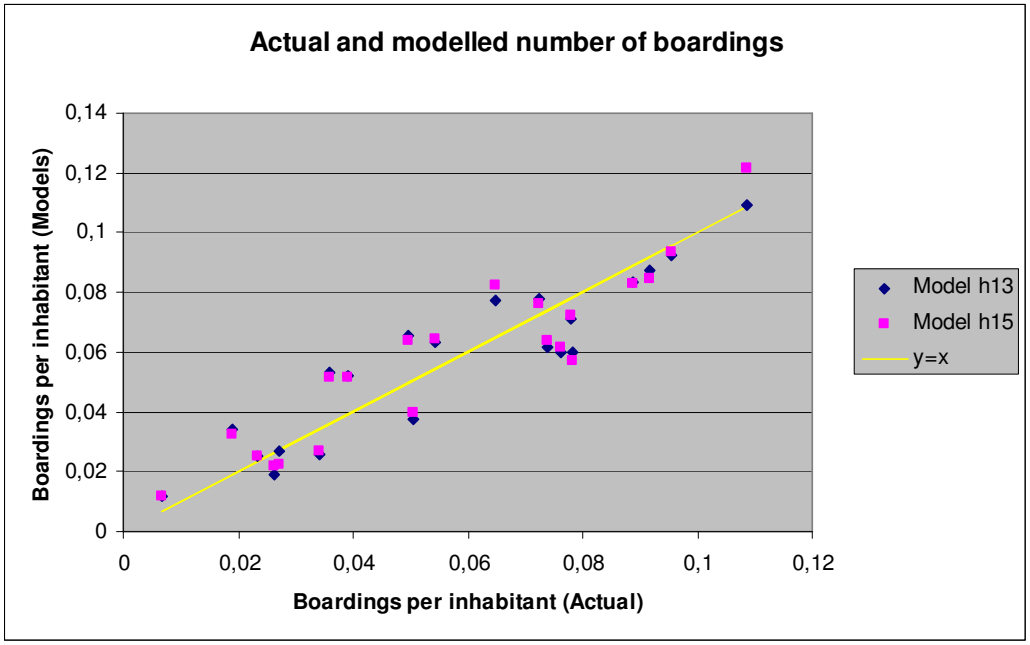


Figure 9.4 Actual number of boardings and model outcomes with the harbour zone included

University and harbour
The last regression models are made including both the (non adjusted) university zone and the harbour zone. The number of college enrolments per inhabitant again is one of the variables included. In table 9.7 the model parameters are displayed.

Model	R^2		Constant	% of people aged 15-24	Frequency	Jobs per inhabitant	Punctuality	Address density	College enrolments per inhabitant
uh13	0.953	Coefficient	-0.09000		0.00103	0.00287	0.00169	-0.00002	0.06077
		Significance	0.017		0.000	0.000	0.007	0.002	0.000
uh15	0.952	Coefficient	-0.09769	0.00106	0.00103	0.00287	0.00168	-0.00002	0.04272
		Significance	0.015	0.430	0.000	0.000	0.008	0.005	0.079

Table 9.7 Model parameters with both the university and harbour included

Just as with the models of the harbour alone the parameter of the percentage of people aged 15-24 is low for model uh15. Model uh13 does have variables with very high significance levels and also explains extremely much of the variance. The reason for this extremely high value of R^2 is again the extreme value of the number of boardings per inhabitant in the university zone (see figure 9.5) and the high significance level is the effect of the extreme value of the number of jobs per inhabitant in the harbour zone. As far as model parameters are concerned only model uh13 is a good model when both zones are taken into account.
In figure 9.5 it can also be seen that for both zones, university and harbour, the estimated number of boardings are very close to the actual number of boardings.

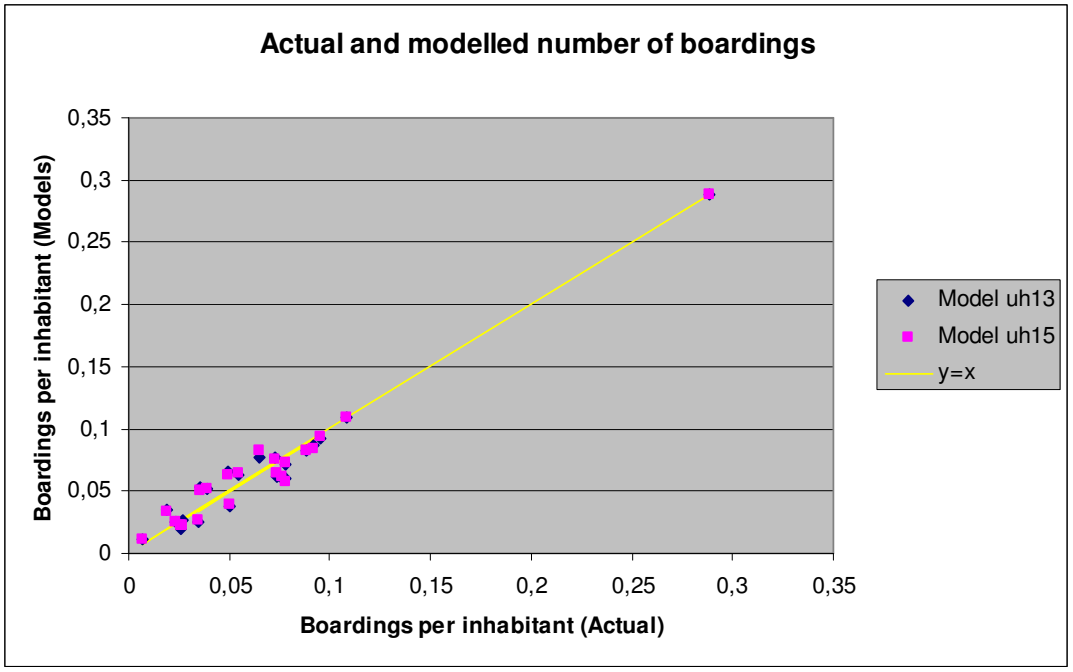


Figure 9.5 Actual number of boardings and model outcomes with the university and harbour zone included

9.6 Conclusion

From the 18 different models that were made excluding the three zones with extreme values two models (13 and 15) are possibly best. These models now have to be validated to check whether or not variables are forgotten or whether or not the model also works for another year. One thing that is already clear is that the frequency plays an important role as determinant of the demand for bus transport.

Also models were estimated with some of the zones with extreme values included. Based on the model parameters these models seem to be better, but in many cases it can be argued whether the models really are better. At this moment it can already be concluded that all models with the number of people aged 15-24 included did not satisfy the requirements on the significance level. Also adjusting the number of boardings per zone in the university zone did not improve the model estimates and model parameters. Because it is desirable to have a model that can be applied on all zones model uh13, with both the university and the harbour zone included, is also validated next.

10 Validation

The different models that are the result of the regression phase now have to be validated. First the outcomes of the models are compared to different variable values to check whether or not the spread can be explained or whether it is purely random.
After that a validation with a new year is performed. This is done with data of the year 2004. First the data are described shortly, because some sources or assumptions are somewhat different. After that the different models are used to determine a number of boardings per inhabitant for the different zones and the numbers are compared to the counted number of boardings and the data are used to estimate 2004 models. Based on this comparison the best model is determined.

10.1 Model outcomes

The outcomes of the three models that represented the dataset of 2008 best (the models 13 and 15 and uh13) are now described in more detail. In this paragraph the residuals are described and the residuals are plotted against different variables, mainly not included in the models, to check whether or not this is a variable that could improve the model. Because the purpose also is to compare the different models with each other, the university and harbour zone, that are not included in the model estimation of model 13 and 15, are not taken into account.

10.1.1 Residual values

In figure 10.1 the residual values for all zones are visualised. The residual value is equal to the modelled value minus the actual value for the number of boardings per inhabitant. When the residual value is positive the model predicts a higher number of boardings than the actual number. The first interesting notice is the fact that in most cases the residual values of the models have the same sign, or are located around zero together. All models hence either overestimate or underestimate the number of boardings per inhabitant or predict the number of boardings quite precise. This can indicate that the models miss a determinant of the demand for bus travel or this is due to randomness in the dataset.

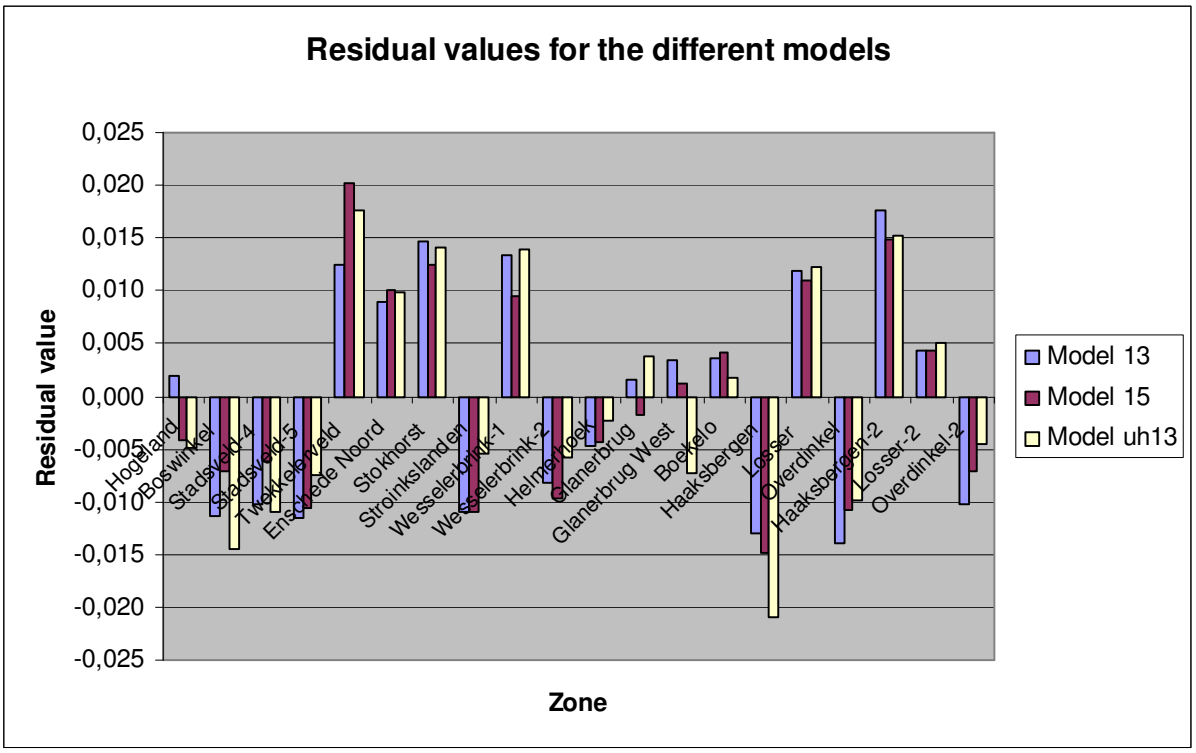


Figure 10.1 Residual values of the different models for all zones

Looking just to the R^2 values of the different models would suggest that model uh13 is the best model. The explained variance however is that high for this model because two extremes are included that

both are optimised by the inclusion of one variable (college enrolments for the university and jobs per inhabitant for the harbour). Therefore it is interesting to look at the explained variance of model uh13 in the other (20) zones. This value can be calculated using formula 10.1.

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} \quad (10.1)$$

With:

$$SS_{err} = \sum_i (y_i - f_i)^2 \quad (10.2)$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \quad (10.3)$$

Where:

$\bar{y}$ = The mean (average) value of the number of boarding per inhabitant

y_i = The actual value of the number of boardings per inhabitant in zone i

f_i = The modelled value of the number of boardings per inhabitant in zone i

In table 10.1 it can be seen that the explained variance of model uh13 is lower than that of the other models. This table also shows the summation of the absolute values of the residuals. The lower these absolute residuals are, the better. It can be seen that the best model is still model 15 followed by model uh13 and model 13.

Model	R^2	Sum of absolute residuals
13	0.798	0.1882
15	0.804	0.1751
uh13	0.778	0.1878

Table 10.1 R-square of models for the different years, without university and harbour zone

10.1.2 Actual number of boardings

In figure 10.2 first the residual values are plotted against the actual number of boardings per inhabitant. Also linear trend-lines are included to check for a relation between the two variables. It can be seen that a slightly downward relation is present between the residual value and the actual number of boardings per stop. In general, the number of boardings per inhabitant is overestimated when the number of boardings is low and underestimated when the number of boardings is high. Ideally the trendlines should follow the x-axis (y=o). It can be seen that model 15 deviates least from the desired line and model uh 13 deviates most. With correlation coefficients of between -0.392 and -0.361, the relations are reasonably obvious. Because the actual number of boardings is unknown when a model is used to estimate future boardings, the actual number of boardings cannot be used to adjust the modelled number of boardings. The modelled number of boardings can also not be used, because the relation between the residuals and the modelled number of boardings is not present because the models are fitted to the data.

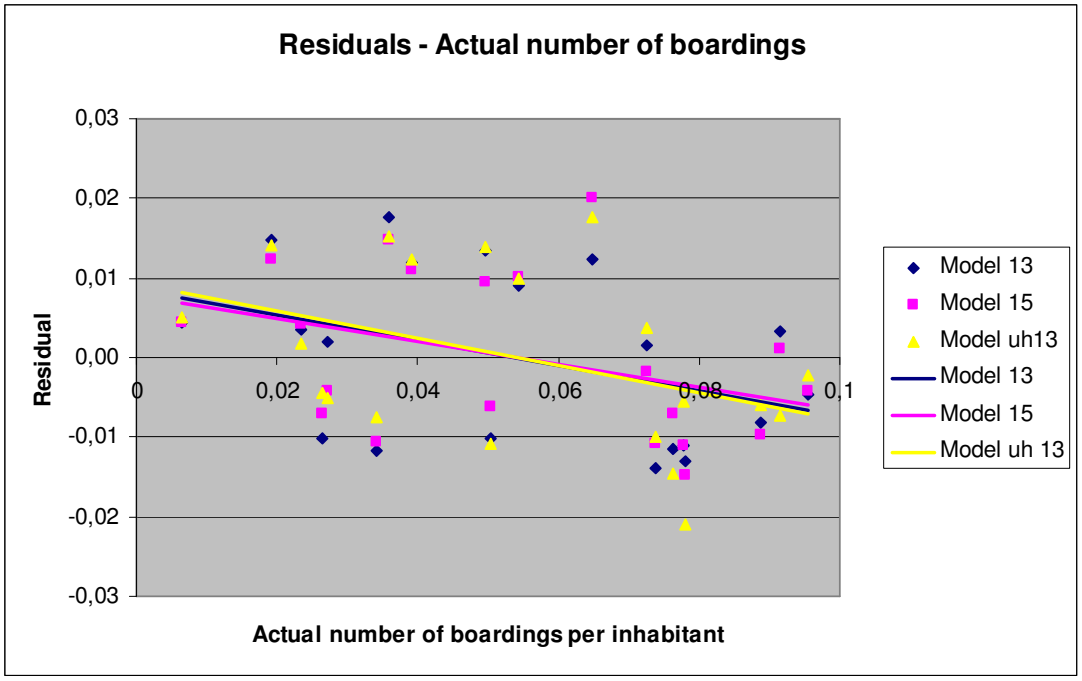


Figure 10.2 Residual values plotted against the actual number of boardings

10.1.3 Frequency

The frequency of services is, as concluded earlier, the most important determinant of the demand for bus transport in an area. It therefore is interesting to look how the residuals are related to the frequency. As could be expected, no relation is found between the residual values and the frequency (figure 10.3. The residuals do not seem to become more spread with a growing frequency. This indicates that no heteroscedasticity is present, which is one of the requirements for doing regression analysis.

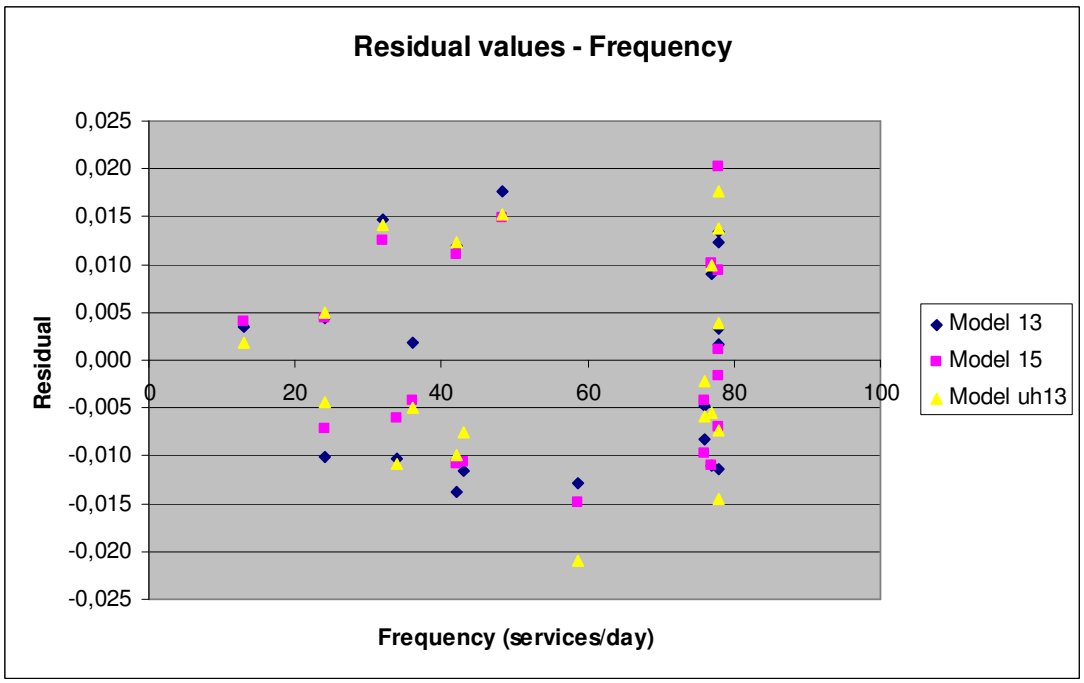


Figure 10.3 Residual values plotted against the frequency

10.1.4 Car ownership

In literature car ownership is mentioned many times as an important determinant of the demand for bus transport. Car ownership however is not a variable that is included in one of the models. To check whether car ownership is related to the residual of the models, and therefore does explain part of those residuals, the variables are plotted against each other in figure 10.4. From this figure it can be concluded that no relation is present between the value of the residual and car ownership. Including this variable hence will not improve the model.

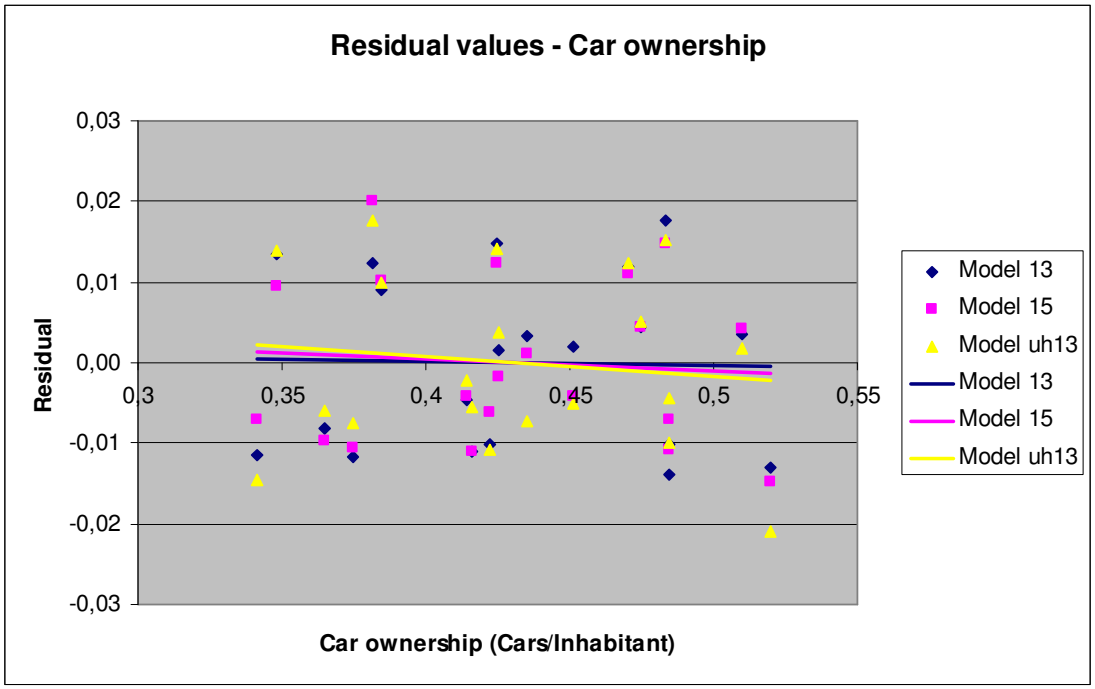


Figure 10.4 Residual value plotted against car ownership

10.1.5 Distance to stop

Another variable that is not included in the models is the average distance to a stop in a zone. In figure 10.5 the average distance towards a stop is plotted against the residuals. Only model 13 gives a slight relation between the distance to stop and the residuals, but the correlation coefficient is still not higher than 0.172, which is too low to assume that a real relation exists. Including the distance to stop hence will not improve the model.

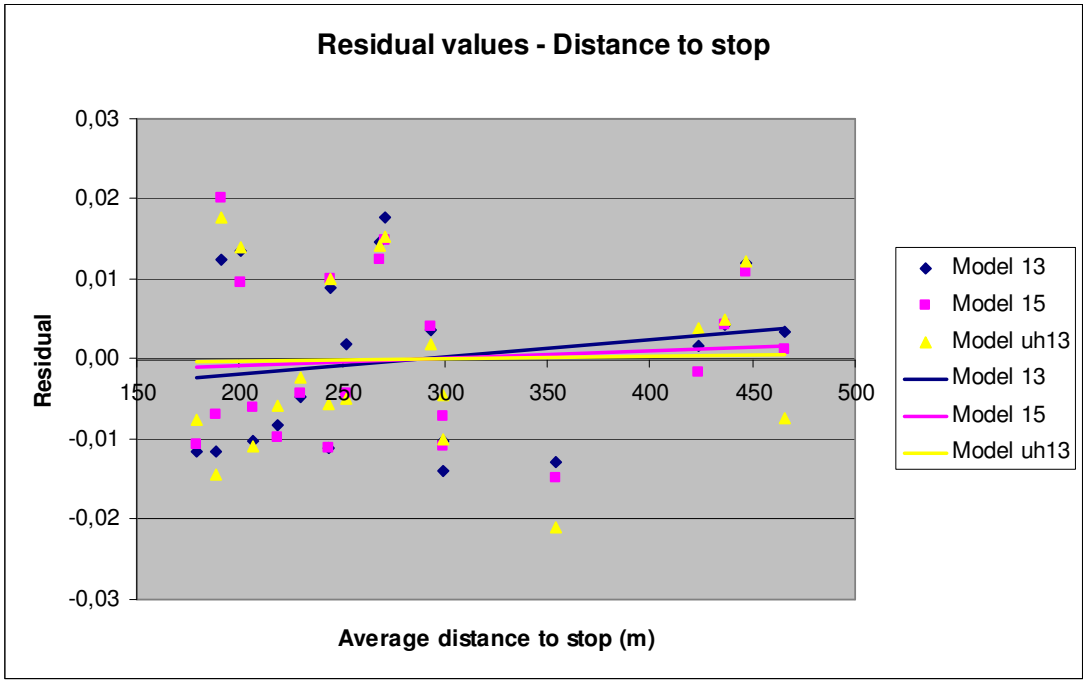


Figure 10.5 Residual values plotted against the average distance to a stop

10.2 Validation with 2004 data

The second validation experiment is to use another year: 2004. The data of this year are used in the models that were calibrated earlier to see whether or not the models also fit for other years in the same study area and 2004 models are estimated to check for different parameter values and significance levels. First the data that are used are described and after that the outcomes of the models are presented.

10.2.1 Data

For many variables the data source is the same or comparable to the source of the 2008 data. The data sources are presented in this paragraph shortly. First however the differences in zones and lines in the research area are described.

Neighbourhoods and lines
The neighbourhood borders in 2004 were slightly different from those in 2008. This mainly is due to the building of new houses. Also in the regional model (RVM) some changes are made on the zonal subdivision. Because of that the union with ArcGIS between the neighbourhoods and the RVM zones is redone. Also some research objects are affected by the border changes of neighbourhoods. Between 2004 and 2008 some routes of lines have been changed. This only accounts for line 2 through the Northern neighbourhoods of Enschede and line 3 through Glanerbrug. For the zonal subdivision this however does not have consequences.

Boardings
To check the models the number of boardings at the different lines should be known. Just as for 2008 NVS counts are available for 2004. Because no valid NVS data were available for one of the zones in Haaksbergen and no SLO data are available of the previous concession period this zone is not taken into account in this validation.

College enrolments
College enrolments were already available for nearly all educational institutes. Only for Saxion Hogescholen the number of enrolments is not known. However, because this institute is located in the centre, and the centre is not included in the model, this is not a problem.

Inhabitants and jobs
The number of inhabitants and jobs in 2004 comes from the regional model of 2004 (RVM 2004). For the university the number of inhabitants is very low compared to the 2008 data (360 in 2004 and 2.600

in 2008). In the Neighbourhood map the number of inhabitants is also much higher (2.450) than in the regional model. The Neighbourhood map data are used to prevent strange outcomes of the model.

Frequency
The average weekday frequency is known from the schedule book (Connexxion, 2003).

Percentage of people aged 15-24
The percentage of people aged 15-24 is known from the Neighbourhood map of 2004 (CBS, 2005).

Punctuality
The SabiMIS reports, from which the punctuality numbers of 2008 come, were introduced in 2006. After some start-up problems the numbers of 2008 are the first reliable punctuality numbers in these reports. For 2004 there hence is no information about the punctuality. Between 204 and 2008 only slight alterations to the routes and infrastructure in the research area were made. Because of that it can be assumed that the punctuality in 2004 did not differed that much from the punctuality in 2008. Because of that the punctuality numbers of 2008 are used as the representation of the punctuality of 2004.

Address density
The address density is known from the Neighbourhood map of 2004 (CBS, 2005).

10.2.2 2004 data in 2008 model

The data form the datasets that are described above and the models are used to estimate the number of boardings per inhabitant in for all zones. In table 10.2 the explained variance of the different models is presented. It can be seen that the explained variance of the 2004 dataset is highest in model 15, which is the model that also had the largest part of the variance explained in 2008. The explained variance of model uh13 is close to that of model 15 and model 13 really performs worse. Overall the R^2-values are lower than for 2008, because just below 65% of the variance is explained, which means that more than 35% is not explained. As the model is fitted for 2008 it is however logical that the explained variance is lower for the 2004 dataset. An explained variance of 64% is still quite high.

Model	2008	2004
13	0.798	0.588
15	0.804	0.640
uh13	0.778[15]	0.632

Table 10.2 R-square of models for the different years, without university and harbour zone

In figure 10.6 the actual and modelled values of the number of boardings per inhabitant are visualised. Interesting is that trendlines through the datapoints of the estimates are less steep than the line representing the perfect fit (y=x). It can be seen that the number of boardings per inhabitants are underestimated by the models when the number of boardings is high. This indicates that larger numbers of boardings in a zone are systematically underestimated. In the 2008 model this also happened, but for the 2004 data the underestimation is larger. The fact that model 15, on average, is closest to the y=x line confirms the fact that the explained variance is highest with this model.

[15] Like in table 10.1 this is the R^2 value for the 20 zones that are also included in the “normal models”

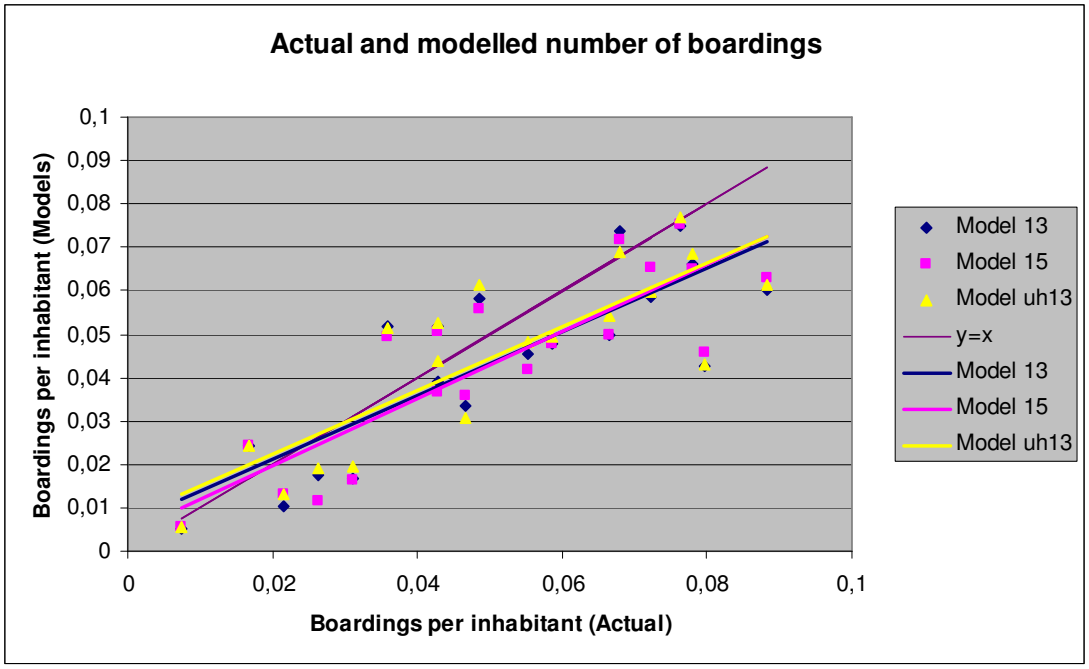


Figure 10.6 Actual and modelled boardings per inhabitant of 2004

Estimating the number of boardings in the university and the harbour zone
For the model with both the university and the harbour zone included (uh13) also the model is used to check whether or not the 2004 numbers of boardings on the two zones not included in the other models is estimated correctly. In table 10.3 the model estimates and the actual number of boardings for the two zones are shown.

Zone	Actual	Estimate	Deviation
University	0.298	0.270	-9%
Harbour	0.059	0.082	+39%

Table 10.3 Actual and estimated number of boardings for the university and harbour zone

The number of boardings in the university zone is estimated quite good. Because of that it can be concluded that the relation between the number of college enrolments, which is an important variable for the estimation in this zone, and the number of boardings per inhabitant hence is quite stable.
In the harbour zone the number of boardings per inhabitant however is substantially overestimated. The relation between the number of boardings and the number of jobs, the most important variable for this zone, hence seems to be less stable than the relation with the number of college enrolments.

10.2.3 Analysis

The deviation of the different (for 2008 fitted) models to the 2004 data is interesting and is now analysed in further detail.

Boarding change
Because of the changes in the service provided by the bus system between 2004 and 2008 it is said that the bus use has increased substantially (30% between 2002 and 2007) in Twente (Heuts, 2009)[16]. The question now is whether this also is the case for the research area part of Twente. In figure 10.7 therefore the number of boardings per inhabitant in 2004 is plotted against the number of boardings in 2008. In this figure all points below the y=x line indicate a decrease of the number boardings between 2004 and 2008 and all points above the line indicate an increase of the number of boardings. It can be seen that for most zones the number of boardings has not changed much. When there is change, for most zones the change is within a margin of 0.02 boardings per inhabitant. For one zone the relative number of boardings per inhabitant did decrease a lot: Stokhorst. This decrease can be explained by

[16] This rise in ridership is accompanied by lower prices, which would be a reason for overestimation of the number of boardings in 2004 rather than an underestimation

the fact that a new “service” line was introduced in the area in the period between 2004 and 2008 that also gives good access to the central train station and city centre.

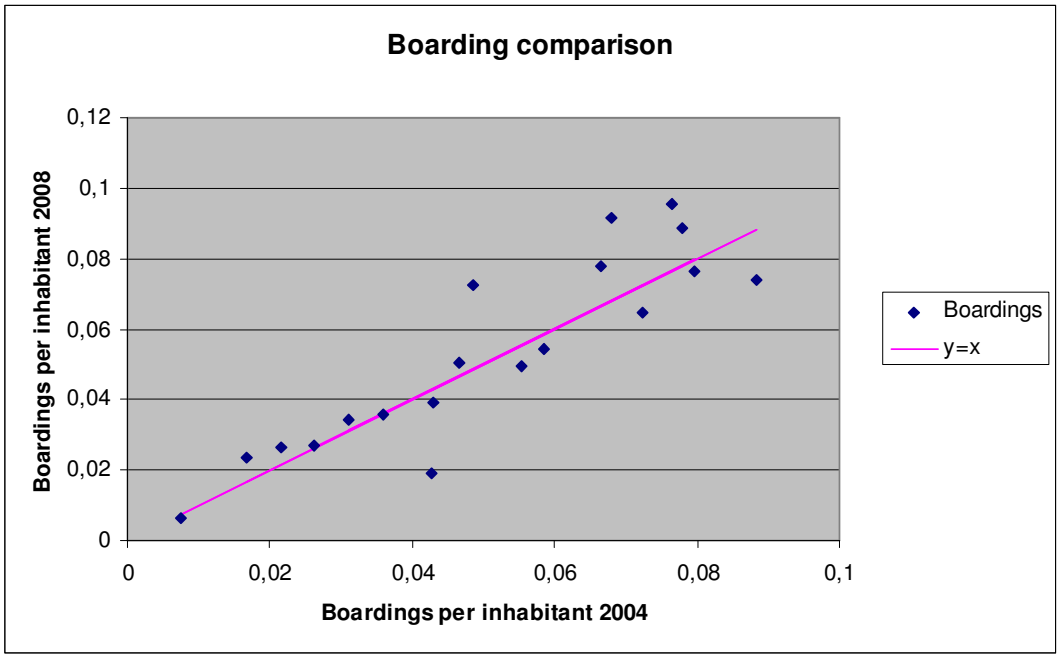


Figure 10.7 Comparison between the number of boardings in 2004 and 2008

Frequency

Because the frequency is the most important determinant and frequency is one of the variables that changed most between 2004 and 2008, this is the first variable to look at to explain the differences between the estimated and actual number of boardings. In figure 10.8 the change in frequency and the change in the number of boardings per inhabitant are plotted against each other. It can be seen that, especially in the zones that have most extra services a day, the change in the number of boardings is very different. Some zones even show a decrease in the number of boardings when the frequency is increased by 16 or 17 services a day. The trend line through the datapoints however is positive and the parameter value for the frequency (0.00092) is comparable to the one found in the model of the number of boardings (which is about 0.001). This indicates that an increase in frequency, on average, leads to larger numbers of boardings. This confirms that frequency is an explanatory variable of the demand for bus transport rather than a variable that depends on the number of boardings[17]. One would however expect that, assumed that all other variables stay equal, the trend line would go through the origin of the plot. Overall the number of boardings, corrected for the frequency, hence seems to be relatively higher in 2004 than in 2008.

[17] When variables are related to each other two possible relations could be present. Variable A depends on the variable value of variable B or variable B depends on the value of variable A. It could hence be that the frequency is high in certain zones because the number of boardings per inhabitant is high in those zones. If that would have been the case an increase in frequency does not lead to an increase in the number of boardings per inhabitant. The model would have been designed wrongly in that case.

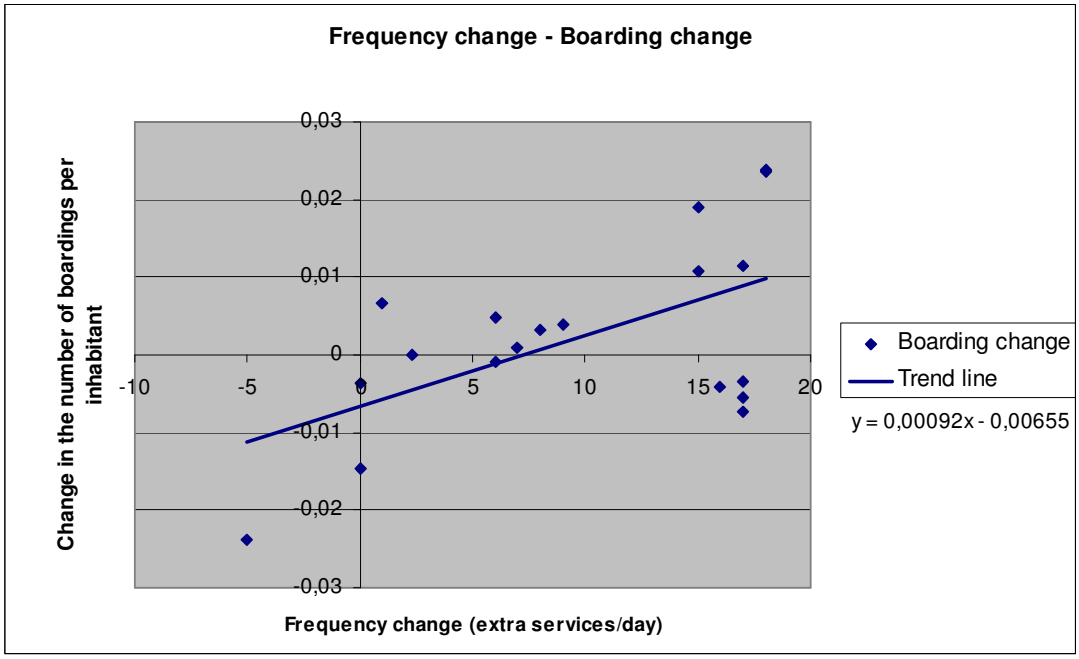


Figure 10.8 The change in the frequency plotted against the change in the number of boardings

The four zones where the frequency increased much and the number of boardings per inhabitants decreased are: Boswinkel (line 1), Twekkelerveld (line 1), Wesselerbrink (line 1) and Enschede North (line 2). For line 2 the decrease can be related to the route change between 2004 and 2008. For two of the other zones a new competitive line was introduced. The frequency on lines 1 and 2 was increased form 4 per hour during daytime in 2004 to six per hour in peak periods and stayed 4 in non-peak periods. This increase did not lead to an increase in boardings. A possible explanation for that is that the number of boardings per inhabitant in these zones already reached its maximum (ceiling) with the previous number of services per day. When a certain, probably zone specific, ceiling is reached increasing the frequency then does not lead to an increase in the number of boardings per inhabitant. When this ceiling exists, it most likely is dependent on the Origin-Destination relations of people travelling from the zone and the destination that can be reached with the line through the zone.

The question now is how it can be explained that the trend line in figure 10.8 does not go through the origin of the plot. There hence is an overall decrease in the number of boardings corrected for the frequency. The most plausible explanation is the nature of the boarding counts in combination with the weather circumstances. The passenger counts are undertaken in two weeks in November each year. The weather in two counting years however can be different. If one period is very cold and wet and the other is warm and dry the number of passengers is different, even if all other variables stay the same. It is possible that in 2004 November was a dry and cold period which made more people to choose for the bus instead of the bicycle. If the weather in November 2008 the weather was better people might have been using the bicycle relatively more, which could have led to lower numbers of passengers (and boardings). In the data centre of the Dutch Meteorological Institute (KNMI)[18] it can be seen that November 2004 indeed was somewhat colder than November 2008 and that the precipitation in 2008 mainly fell during the night. These data hence increase the suspicion that the weather plays a role in the difference in the underestimation of the 2004 boarding numbers. When this is the case all data points in figure 10.8 should be transposed along the y-axis, which means that the trend line comes closer to the origin.
A suggestion to improve the models is to fit models on as many years as possible of which some will have weather circumstances that are increasing bus use and some will have weather circumstances that decrease bus use. Together the model then gives an estimation of the average number of boardings per inhabitant in average weather circumstances. Because of time constraints this research does not make this extra step.

[18] http://www.knmi.nl/klimatologie/maand_en_seizoensoverzichten/index.html

10.2.4 2004 models

To check whether the change in explained variance can be attributed to certain variables the three models are also estimated based on the 2004 data. When an estimated parameter value is (very) different in the 2004 models than in the 2008 models, the influence on the number of boardings per inhabitant is not stable over time, which can explain the change in explained variance. In this way it can also be seen whether or not the parameter values are also significant for 2004. The model parameters of the 2004 models are shown in table 10.3. This time the explained variance of model 15 is much higher than those of the other two models.

Year	Model	R^2		Constant	% of people aged 15-24	Frequency	Jobs per inhabitant	Punctuality	Address density	College enrolments per inhabitant
2004	13	0.693	Coefficient	-0.08321		0.00103	0.00126	0.00148	-8.9E-06	
			Significance	0.041		0.000	0.958	0.025	0.143	
2008		0.797	Coefficient	-0.10171		0.00101	0.01792	0.00183	-1.9E-05	
			Significance	0.010		0.000	0.185	0.004	0.002	
2004	15	0.815	Coefficient	-0.12524	0.00423	0.00098	0.01145	0.00156	-2.1E-05	
			Significance	0.002	0.007	0.000	0.542	0.005	0.003	
2008		0.82	Coefficient	-0.11670	0.00161	0.00099	0.02293	0.00185	-2.5E-05	
			Significance	0.006	0.246	0.000	0.108	0.004	0.003	
2004	uh13	0.952 (0.672)[19]	Coefficient	-0.08069		0.00105	0.00187	0.00144	-9.0E-06	0.07442
			Significance	0.038		0.000	0.010	0.025	0.127	0.000
2008		0.953 (0.778)	Coefficient	-0.09000		0.00103	0.00287	0.00169	-1.9E-05	0.06077
			Significance	0.017		0.000	0.000	0.007	0.002	0.000

Table 10.3 Model parameters for the 2004 and 2008 models

In the table it can be seen that the frequency parameters did not changed that much over the years. The earlier noted underestimation of high frequency zones hence is not caused by the change in the relation between the number of boardings per inhabitant and the frequency, but in the change of other relations. The punctuality parameters also did not change much, as did the values of the constants. The most remarkable parameter change happens with the number of jobs per inhabitant. In 2008 this parameter has much higher values than in 2004. Because the parameter values are lower, less certainty is whether this value significantly differs from zero, which makes that the significance level does not satisfy the requirements for models 13 and 15. The fact that the parameter value is also lower in 2004 than in 2008 for model uh13, explains the overestimation by the 2008 model of the 2004 number of boardings per inhabitant in the harbour zone.
In figure 10.9 a plot is shown of the 2004 and 2008 number of boardings per inhabitant and the corresponding number of jobs per inhabitant (the university and harbour zones are not included). Even a slight downward slope can be discovered for 2004, while the parameters in the model are positive. It can however also be seen that this relation does not have a high significance level because the points are pretty spread and the line is nearly horizontal, so not much different from zero. For 2008 the slope is upward and the significance level again is not very high. For both years the location of the number of zones with high numbers of jobs per inhabitant seem to determine whether or not the relation is clearly positive or not. When these zones are excluded no clear trend can be observed.

[19] Between parentheses the adjusted R^2 when only the zones included in the “normal” models are taken into account

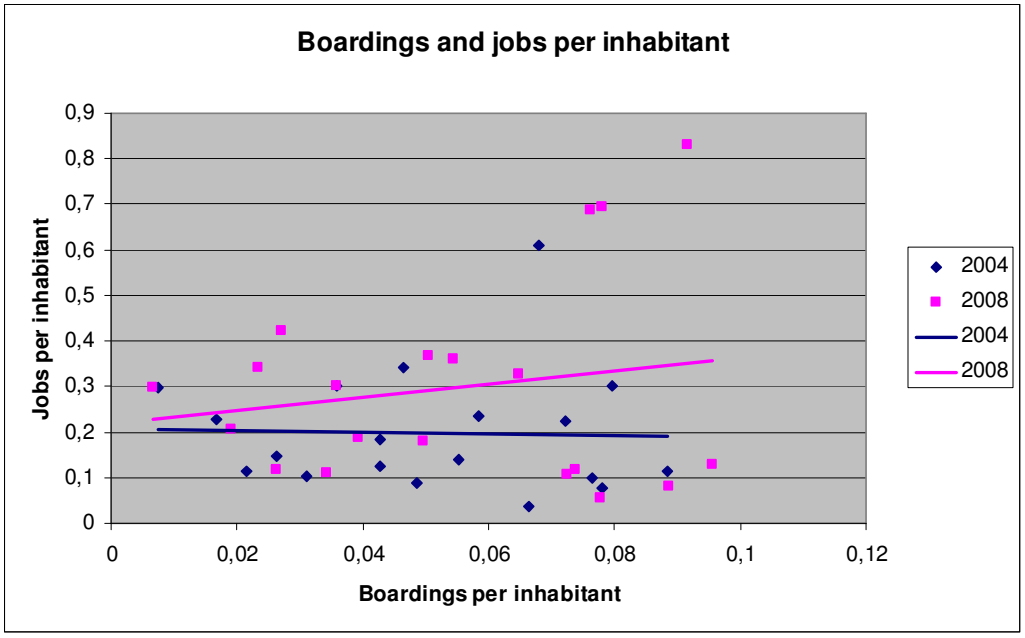


Figure 10.9 The number of boardings and jobs plotted against each other for the two years

For the address density two of the three models (13 and uh13) show a substantial change in the parameter (from -19E-6 in 2008 to -9E-6 in 2004). In 2004 the effect of the address density hence is less negative than in 2008. The slight relation also makes that the significance level is not that high for the variable. The increase in the negative value of the parameter between 2004 and 2008 indicates that the increase in the number of boardings per inhabitant was larger in less dense, so less urbanised, zones than in dense zones in the centre of Enschede.
The number of people aged 15-24 (only included in model 15) in 2004 has a higher parameter and significance value. The influence of students in 2004 hence is higher than in 2008. This can also be concluded because of the higher parameter value for the number of college enrolments.

10.3 Conclusion

From the model outcomes it can be concluded that none of the studied variables has a clear relation with the residual values of the model estimates. However all three models do overestimate small numbers of boardings and underestimate large numbers of boardings. Based on this validation model 15 is to be preferred because the explained variance is highest and the correlation with the non-included variables is lowest.
From the 2004 dataset the conclusion is that the models (both with and without the university and harbour) can predict the number of boardings in 2004 with a reasonably high precision (nearly 65% for models 15 and uh13), but that the models underestimate boarding numbers in 2004 in high frequency zones more than in 2008. The most plausible explanation for this is different weather circumstances for the two counting periods and the introduction of a new line. It is also possible that there is a maximum number of boardings per inhabitants for each zone that makes that frequency increase sometimes does not lead to higher numbers of boardings per inhabitant. When models are estimated based on the 2004 data also model 15 performs best as far as explained variance is considered. The significance level of the number of jobs per inhabitant however does not satisfy the requirements, but the significance level of the percentage of people aged 15-24 does.

Considering the different validation outcomes model 15 is the best model. The disadvantage of this model is however that it cannot be used for zones with only business areas and for the university zone. A valuable model that includes these zones is however not at hand at the moment, because too less zones of this kind are present at the research area which makes the values of some parameters sole standing outliers.

11 Conclusion

The objective of this research was to develop a model that could be used to estimate the bus demand in three municipalities in the area of Twente in the Netherlands. This objective was translated in different research questions. In this concluding chapter those questions are answered. When those questions are answered it is possible to determine the best model for the demand for bus transport. This model is described in the second part of this chapter. After that the relations and the variables included and not included are compared to literature and differences are explained as far as possible. The chapter ends with some recommendations for further research.

11.1 Answering research questions

In paragraph 2.3 six research questions were formulated to be answered during the research. All of them are answered here.

11.1.1 What determines the demand for bus transport? (Theory)

In literature many determinants of the demand for bus transport are described. They can be subdivided into four categories: socio economical characteristics, built environment characteristics, network characteristics and personal preferences. In appendix A an overview of many of those variables and their interdependencies is displayed.

11.1.2 What are the most important variables that need to be included in the model?

For all categories of possible determinants, except the personal preferences, a selection was made of the variables that were classified as very important determinants of the demand for bus transport by certain researchers or variables that could be important because of specific Dutch circumstances (e.g. students that travel for free).
The variables included are[20]:

- Socio economic
  - Car ownership
  - Business car ownership
  - Income
  - The percentage of students (the percentage of people aged 15-24)
- Built environment
  - Inhabitants
  - Jobs
  - Accessibility of destinations (Relative job accessibility by car or bicycle, travel time to the city centre by bus or relative travel time towards the city centre by car or bicycle)
  - Spatial density (population, job and address density)
  - College enrolments
- Network
  - Frequency
  - Punctuality
  - Distance to stop
  - Train station

11.1.3 How can the research area be subdivided into research objects?

The research area is subdivided into zones (research objects) that are formed by groups of neighbourhoods. This scale level is used because the boarding counts can only be attributed to groups of neighbourhoods rather than to neighbourhoods alone. In the end the research area is subdivided into 23 zones, of which some overlap each other because some neighbourhoods are serviced by more than one line with different destinations/routes.
Because many of the variables have influence on people's individual choices, the dependent variable is the number of boardings per inhabitant rather than the number of boardings in a zone itself.
Because most data are available for 2008, this is the research year for the study.

[20] Between parentheses different measures for the variable

11.1.4 What data are needed?
In chapter 7 all data that are used for the research are described. Some data were accessible easily, while others had to undergo a large number of calculating or computing steps before they could be used in the research.

11.1.5 What are the quantitative relations?
With the computer programme SPSS regression was undertaken (chapter 9) to estimate different models that described the relation between the dependent variable (the number of boardings per inhabitant) and a limited number of independent variables. The basis for this regression was the correlation analysis of chapter 8, which was used to determine what variables could and could not be used together in one model.
In this regression analysis three possible models that had a high explanatory value and satisfy the requirements regarding parameter significance were estimated.

11.1.6 How can the model be validated?
The models following from the regression analysis are validated in chapter 10. The validation is done using the values of (non) used variables and another year (2004). From the validation the following conclusions were drawn:

- Model 15, with independent variables: frequency, punctuality, percentage of people aged 15-25 and the number of jobs per inhabitant performs best for the 2008 data.
- All models estimate the 2004 numbers of boardings worse than they fit to the 2008 data.
- There seems to be a certain, zone specific, ceiling for the number of boardings per inhabitant. Increasing the frequency when this ceiling is reached therefore does not lead to an increase in the number of boardings per inhabitant.
- The nature of the boarding counts together with varying weather circumstances are the most plausible explanation for this fact.
- Model 15 also is best for the 2004 data.
- Model 15 is used, although no business and university zones can be included in this model

11.2 Definitive model
In this paragraph the best model is described and some important notices are made that need to be taken into account when the model is used.

11.2.1 Model description
When all zones but the university and the harbour (and the city centre) are taken into account the best model is the model (15) with the parameters that are presented in table 11.1. It can be seen that all parameter values are significantly different from zero with 85% certainty. The value of R^2 is 0.804, which means that about 80% of the variance is explained by the model. For 2004 the variance explained is about 64%.

Model 15 (R^2 = 0.804)	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta	B	Std. Error
Constant	-0.117	0.036		-3.230	0.006
% of people aged 15-24	0.161	0.133	0.183	1.212	0.246
Address density	-2.46E-05	6.70E-06	-0.593	-3.669	0.003
Frequency	0.00099	0.00013	0.852	7.869	0.000
Jobs per inhabitant	0.0229	0.0134	0.188	1.716	0.108
Punctuality	0.00185	0.00054	0.411	3.441	0.004

Table 11.1 Model parameters for the definitive model

Constant
The value of the constant is -0.117. This means that the other variables have to have some base value in order to get a positive number of boardings (per inhabitant).

Percentage of people aged 15-24
The parameter value for the percentage of people has a value of 0.161. When the percentage of people aged 15-24 in a zone increases from 10% to 20%, the number of boardings per inhabitant increases with 0.0161. For a zone with 1000 inhabitants this means an increase of 16.1 boardings a

weekday. The range of the values is between 8.8% and 17.9%. The maximum difference in boardings per inhabitant then is 0.0147.

Address density
The parameter value for the address density is -0.0000246. This value is much lower than the one for the percentage of people aged 15-24 because the value of the variables is very different. Address density for instance ranges from 283 to 2439 addresses within a radius of 1 kilometre, while the proportion of people aged 15-24 ranges from 0.088 to 0.179. When the address density is increased with 1000, the number of boardings per inhabitant decreases with 0.0246, which is 24.6 boardings in a zone with 1000 inhabitants. The maximum difference in the number of boardings per inhabitant is -0.0530.

Frequency
The frequency has a parameter value of 0.00099. This means that an increase of the frequency by 24 services a day (one per hour) leads to an increase of the number of boardings per inhabitant of 0.0238 and an increase of the number boardings of 23.8 in a zone of 1000 inhabitants. Because the range of the frequency is 13 to 78 services a day, the largest difference in the number of boardings per inhabitant is 0.0644.

Jobs per inhabitant
The parameter of the number of jobs per inhabitant has a value of 0.0229. This variable has values in the range of 0.056 to 0.830. The difference between the number of boardings per inhabitant in the zone with the highest and the lowest value therefore is 0.0177.

Punctuality
The parameter value of the punctuality is 0.00185. Because the range of the punctuality of the different lines is not very wide (54.988-78.150), the maximum difference in the number of boardings per inhabitant are also not very high (0.0428). This maximum difference is however larger than the maximum difference following from the number of jobs per inhabitant, which indicates that punctuality is relatively more important.

Relative importance
The relative importance of the variables is shown in the fourth column. When a value of Beta is twice as high as another Beta the variable is twice as important. In table 11.1 it can be seen that the frequency is most important. The importance of the address density is about 70% of the importance of the frequency and the other variables are 20% (students and jobs) and 50% (punctuality) as important as the frequency.

11.2.2 Important notices
The way the model was estimated and the outcomes of some validation steps should be kept in mind when the model in used.
The model is estimated based on data of 2008. This means that the model in general performs worse for years other than 2008. This was also shown in the validation with the data of 2004. The performance of the models are however likely to be smaller when changes of the variable values are not that large between years, for instance because both years are close to each other and in the same concession period.
As mentioned in the validation, weather circumstances can play a role in the differences of the number of boardings. The way of counting passengers/boardings and the time period in which this is done therefore are important.
In the validation it is also mentioned that it is possible that zone specific ceilings for the number of boardings per inhabitant exist. When this is the case increasing the frequency does not lead to increased numbers of boardings. When the model is used to estimate the effects of a frequency increase in a zone that (nearly) reached its ceiling the model will overestimate the number of extra boardings.

11.3 Comparing to literature
In this paragraph the parameter values of the different variables that are included in the models are compared to the literature and the reasons that some variables are not included in the models are discussed based on literature.

11.3.1 Variables included

In total five variables are included in the two models. They are all discussed here starting with the most important and finishing with the least important variable.

Frequency

The frequency is the most important variable in the model. Other studies (like: Paulley et al, 2006) also show that the frequency, or the vehicle kilometres, which is directly related to the frequency, is an important variable of the demand for bus transport. It is obvious that a higher frequency leads to higher numbers of travellers/boardings.

Address density

The address density has a negative parameter value which means that the number of boardings per inhabitant is higher in zones that have a lower address density. The study of Balcombe et al. (2004) however shows that people travel by bus more and more often when they live in densely built areas. In the research area the zones that have low address density are mainly located further from the centre of Enschede. In figure 7.8 it can be seen that many of these zones also have high numbers of boardings per inhabitant. A possible explanation can lie in the presence of a train station in the centre of Enschede. People living close to the centre of Enschede hence have the best possibilities to travel by train. When they travel by train they can access without using a motorised mode for access-transport. Cycling and walking for them is a good alternative access mode for car or bus. Therefore bus use is relatively low in these areas. In the zones further from the centre people are more dependent on the bus to reach the train or the train is not an interesting mode for travel at all. Bus use in these zones, like Haaksbergen, Overdinkel, Losser and the outskirts of Enschede, therefore is relatively high. Because these zones all have a relatively low address density, the number of boardings per inhabitant is negatively correlated to address density, leading to a negative parameter value in the model.
When whole public transport zones (like Twente as a whole) are taken into account it is however likely that people in more densely built areas travel by bus more. The fact that operating bus transport is more attractive there will hence lead to higher frequencies, which was found to be the most important determinant of the demand for bus transport.

Punctuality

Punctuality is also an important variable in the models, of which the effects were not studied earlier. According to the models a higher punctuality leads to higher passenger numbers. This agrees with common sense, because people will not use a bus that is structurally early or late.

Number of jobs per inhabitant

The more jobs there are located in a zone the higher the number of boardings per inhabitant. This is in line with a study by Maat & Timmermans (2009).

Percentage of people aged 15-24

Because students travel for free (on certain days) by public transport in the Netherlands it is no surprise that the parameter for the percentage of people aged 15-24 has a positive sign.

11.3.2 Variables not included

Here the variables that are not included in the models, but were included in the research are treated.

Socio economic

Car ownership and business car ownership did not show significant relations with the number of boardings per inhabitant. Many studies (like: Van Acker & Witlox, 2010, Peulley et al., 2006 &Holmgren, 2007) however show that, especially car ownership is one of the most important determinants of the demand for bus transport. The problem with the car ownership data in the Netherlands however is the way they are registered. The data can be registered by a person, a company or a lease company, that all are located on different locations. On average, this leads to an overestimation of car ownership in business areas and an underestimation of car ownership in residential areas. On a municipal level this however probably does not lead to great bias, because the car ownership in the whole municipality is about the average of business and residential zones, but it can lead to larger bias at a smaller level. Because the registered data are the only easily accessible source of data for car ownership and municipalities are split for the research, this problem of assigning

cars to neighbourhoods can be the explanation for the non-significant relations between car ownership and the number of boardings per inhabitant.
Also the income, mentioned by Souche (2010), Holmgren (2007), Bresson et al. (2003) and Fitzroy & Smith (1998), did not show significant relations. Interesting is that the mentioned studies did not agree on whether or not the relations were significant. One of the problems with the data used for this research are the unreliability of the numbers (errors of more than €3.000 can exist in small areas), and the absence of numbers for the rural municipalities of the used year. This can be a reason for the lack of significance. It however also is possible that there simply is no significant relation between the variables.

Build environment
The accessibility of destinations was not included in the model. Only the relative car accessibility showed significant parameters, but the problems regarding the variables background, the unexpected sign and the large effort that had to be done to obtain the variable values led to the exclusion of the model from the models. The other variables were no good representation for the attractiveness of using bus transport.
The number of college enrolments is not included in the model because the university zone was not included. When it is included it does lead to higher numbers of bus boardings, which is according to expectations.

Network
The distance to stop is not included in the models because of the significance level of its parameter. Distance to stop is related to the access and egress times for using the bus. Paulley et al. (2006) mention that larger access and egress times increase the total travel time, which makes the bus less attractive compared to other modes. Because the research area is pretty densely covered with bus stops, the value for the distance to stop did not really differ that much in the different zones. Only in the rural areas the distances were larger, but these were the zones that also had relatively many boardings per inhabitant given the frequency offered there. The negative relation therefore is not found in the research.
The train station dummy variable was not included in the models because the zone in which the train station was located also was not included. No conclusions therefore can be made. This is however an interesting variable for further research.

11.4 Recommendations & Further research
Because of the different (financial and time) constraints of this research and the availability of data there are possibilities to improve the research or to make it more detailed. The recommendations are described in this paragraph.

The numbers of boardings that are used in the research come from counts that are undertaken once every year in November. Because the counts cannot take too much time (the bus driver has to do them while staying on schedule) the number of passengers is only counted on a small number of stops (maximum 4 per direction per line). The consequence for this research is that the neighbourhoods had to be grouped into a relatively small number of zones to have research objects. However in the future reliable at-stop numbers of boardings will become available from the PT-chipcard. When these are available, and all people use this card, research can be done with the bus stops rather than groups of neighbourhood are research objects. Doing so would lead to a lot more cases and a better representation of the local differences. With the research objects used it is for instance possible that in one part of a zone high income people live that do not use the bus and in the other part of the zone low income people that do use the bus. This can lead to a reasonably high income on the zone whereas also many (poor) people use the bus. Using smaller (more detailed) research objects will probably improve the quality of the model and increase the significance levels of parameter values.
Another problem that was encountered was the underestimation of the number of boardings per inhabitant for the validation year 2004. This problem was most likely caused by the NVS count nature and differences in weather circumstances. To correct for these differences it is possibile to fit a model for different years together. In this way the counts of different years with different weather circumstances are taken together and a model is estimated that fits to all years. In this way also the problem of the low number of research objects becomes smaller because the number of zones increases by about 20 with every year that is added. The question however is whether it is wise to do the regression all over again, or whether the limited number of models resulting from the regression

analysis should only be taken into account in such a research. When time is not a problem, redoing the regression analysis probably gives the best models. For some years however some variable values are not available.

All measures of the accessibility of destinations have been found to fail to explain differences in the demand for bus transport with a satisfying significance level. It seems that the used measures were not really good measures for the attractiveness of using a bus from a certain origin. This attractiveness however is still thought to play a major role in the modal choice of people. Further research therefore should be focussed on the computation of a good measure for the attractiveness of using the bus from a zone. Possible measures can be: the relative job/inhabitant accessibility by bus or the relative travel time towards (a) major attraction point(s) by bus. In the first case a model should be used to determine the number of jobs/inhabitants that can be accessed within reasonable time by bus rather than by public transport as a whole. In the second case more precise estimates should be made on the travel time by the different modes than during this study.

In the research area one zone is included that is the destination for many of the trips, the city centre of Enschede which includes a train station. Therefore a dummy variable for a train station was introduced. Because of the extreme number of boardings per inhabitant, the central zone was not taken into account during the regression phase. This variable therefore was not included in one of the many models that were estimated. Further research should be done on the possibilities to include such a zone and whether or not the train station dummy variable should be used or whether the (residential) zones the lines from the centre go to are more important for such a zone.

In the research methodology some possible determinants of the demand for bus transport were chosen to de studied. Of course a lot variables more could have been studied. Variables that can be thought of are: Land use variables (differentiating between industrial, shopping and residential areas), trip purpose variables (differentiating between the trip purpose, like home-work/education and leisure and social), personal preferences and household composition.

Bibliography

Acker, V. van & Witlox, F. (2010), Car ownership as a mediating variable in car travel behaviour research using a structural equation modelling approach to identify its dual relationship. Journal of Transport Geography 18, 65-74

Aizcorbe, A. & Starr-McCluer, M. (1997), Vehicle ownership, purchase and leasing: consumer survey data. Monthly Labor Review, June 1997, 34-40

Balcombe, R., Mackett, R., Paulley, N., Preston, J., Shires, J., Titheridge, H., Wardman, M. & White, P. (2004), The demand for public transport: a practical guide. TRL report TRL593, TRL Limited, Wokingham, Berkshire, United Kingdom

Berkum, E.C. van (2009), PT-systems and networks, presentation for the Public Transport course 22-09-09. University of Twente, Enschede

Bresson, G., Dargay, J., Madre, J.-L., & Pirotte, A. (2003), The main determinants of the demand for public transport: a comparative analysis of England and France using shrinkage estimators. Transportation Research Part A 37, 605-627.

Ceder, A., & Wilson, N. H. (1986), Bus network design. Transportation Research Part B, Vol. 20B, No. 4, 331-344.

CBS (2005), Digitale Wijk- en Buurtkaart 2004. Centraal Bureau voor de Statistiek / Topografische Dienst Kadaster

CBS (2009), Digitale Wijk- en Buurtkaart 2008. Centraal Bureau voor de Statistiek / Topografische Dienst Kadaster

Connexxion (2003), Dienstregeling 2004 Regio Twente. Connexxion

Connexxion (2008), NVS 2008 Najaar connexxion, Connexxion, The Netherlands

Connexxion (2010), MIPOV 2008 (1.0.0.553) April 2010: Concept version, Concession team Twente, Connexxion, Enschede

Erdem, Ö & Houben, J. (2008), De arbeidsmarktpositie van werknemers in 2006, Een onderzoek naar de verschillen in beloning en mobiliteit tussen groepen werknemers. Arbeidsinspectie, Den Haag

Fitzroy, F., & Smith, I. (1998), Public transport demand in Freiburg: why did patronage double in a decade? Transport Policy 5, 163-173.

Friso, K. & Honing, R. van der (2006), Regionaal Verkeersmodel Twente. Goudappel Coffeng, Deventer

Gardner, B. & Abraham, C. (2007), What drives car use? A grounded theory analysis of commuters' reasons for driving, Transportation Research Part F 10, 187-200

Gautier, P.A., Zenou, Y. (2010), Car ownership and the labor market of ethnic minorities. Journal of Urban Economics (2010), doi:10.1016/j.jue.2009.11.005

Geurs, K.T. & van Wee, B. (2004), Accessibility evaluation of land-use and transport strategies: review and research directions. Journal of Transport Geography 12, 127–140.

Haaijer, M.E. (1999), Modelling conjoint choice experiments with the probit model. Proefschrift Rijksuniversiteit Groningen, Labyrinth Publications, Capelle a/d IJssel, The Netherlands

Heuts, P. (2009). Reizigers nemen massaal Twents Ros, Twente geeft voorrang aan Hoogwaardig openbaar vervoer, ov magazine. 05-02-2009, 18-20, retrieved: 20-07-2010 in http://www.keypointonline.nl/PDF/Reizigers%20nemen%20massaal%20Twents%20Ros.pdf

Holmgren, J. (2007). Meta-analysis of public transport demand. Transportation Research Part A 41, 1021-1035.

Jong, G. de, Fox, J., Pieters, M., Daly, A.J. & Smith, R. (2004), A comparison of car ownership models. Transport Reviews 24(4), 379-408

Kuby, M., Barranda, A. & Upchurch C. (2004), Factors influencing light-rail station boardings in the United States. Transportation Research Part A 38, 223-247

Maat, K. & Timmermans, H.P.J. (2009), Influence of the residential and work environment on car use in dual-earner households. Transportation Research Part A 43, 654-664

Mobycon (2006), Halteplan Twente: Eindrapportage, Mobycon, Delft, The Netherlands

Ortúzar, J. de Dios & Willumsen, L.G. (2001), Modelling Transport: Third Edition, John Wiley & Sons Ltd, Chichester, Engeland

Paulley, N., Balcombe, R., Mackett, R., Titheridge, H., Preston, J., Wardman, M., et al. (2006), The demand for public transport: The effects of fares, quality of service, income and car ownership. Transport Policy 13, 295-306

Recker, W.W. (2001), A bridge between travel demand modelling and activity-based travel analysis. Transportation research Part B 35, 481-506

Rijkswaterstaat (2008), Mobiliteitsonderzoek Nederland 2008: Tabellenboek. retrieved: 11-02-2010 in http://www.rws.nl/images/Mobiliteitsonderzoek Nederland 2008 - Tabellenboek_tcm174-277213.pdf

Rye, T., & Scotney, D. (2004), The factors influencing future concessionaire bus patronage in Scotland and their implications for elsewhere. Transport Policy 11, 133-140.

Souche, S. (2010), Measuring the structural determinants for urban travel demand. Transport Policy, doi:10.1016/j.tranpol.2009.12.003.

Stichting Twente Index (2009), Twente Index 2009: Een vergelijkend economisch jaarbeeld van Twente. Stichting Twente Index, Enschede

Taylor, J.R. (1997). An introduction to Error Analysis, 2nd edition. Sausalito (CA), USA: University Science Books, 166-167 (http://books.google.nl/books?id=giFQcZub80oC&dq=an+introduction+to+error+analysis&printsec=frontcover&source=bn&hl=nl&ei=x-HCS7-uLJj4mwPww9jaBg&sa=X&oi=book_result&ct=result&resnum=4&ved=0CCEQ6AEwAw#v=onepage&q&f=false)

Twents (2009), Dienstregeling 2010: Regio Twente. Connexxion, 2009

Appendix A: Variables and their interdependencies

In this appendix an overview of the variables that are related to the demand for bus transport and their mutual dependencies is given. Also indirect determinants are included. An explanation of the figures can be found on page 90.

Socio economical

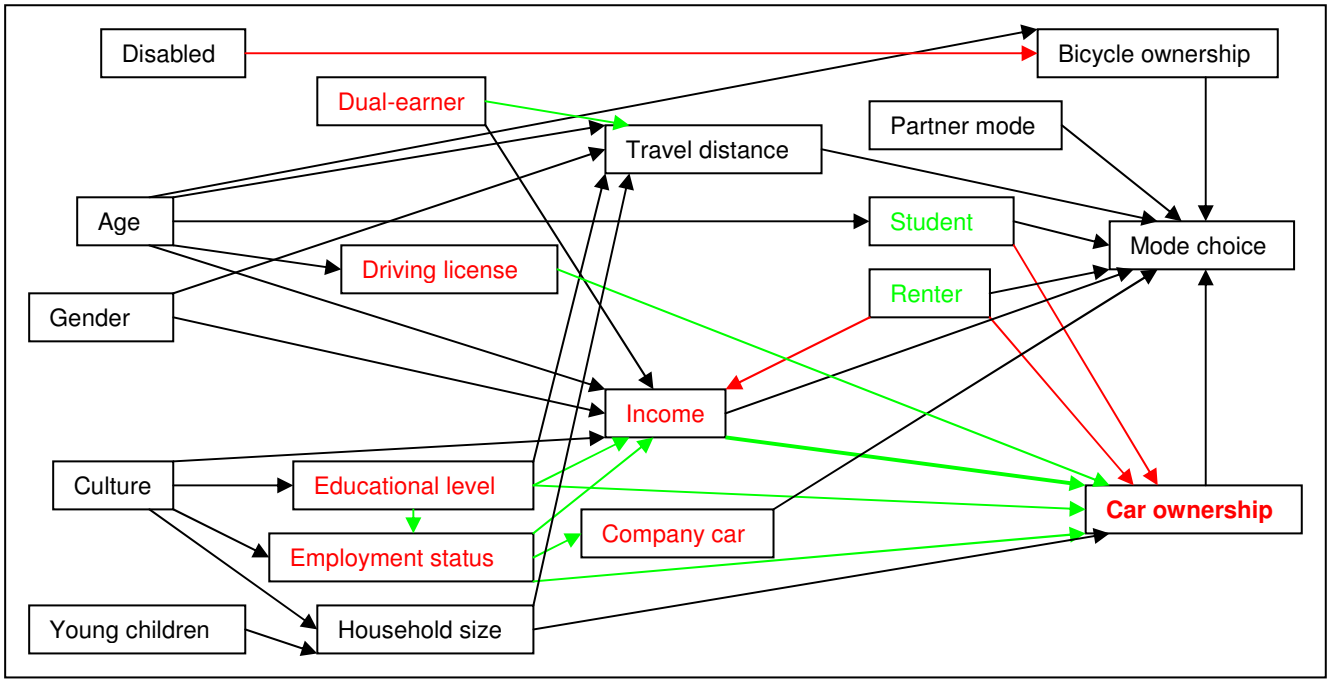


Figure A.1 Socio economical variables related to the demand of bus transport

Built environment

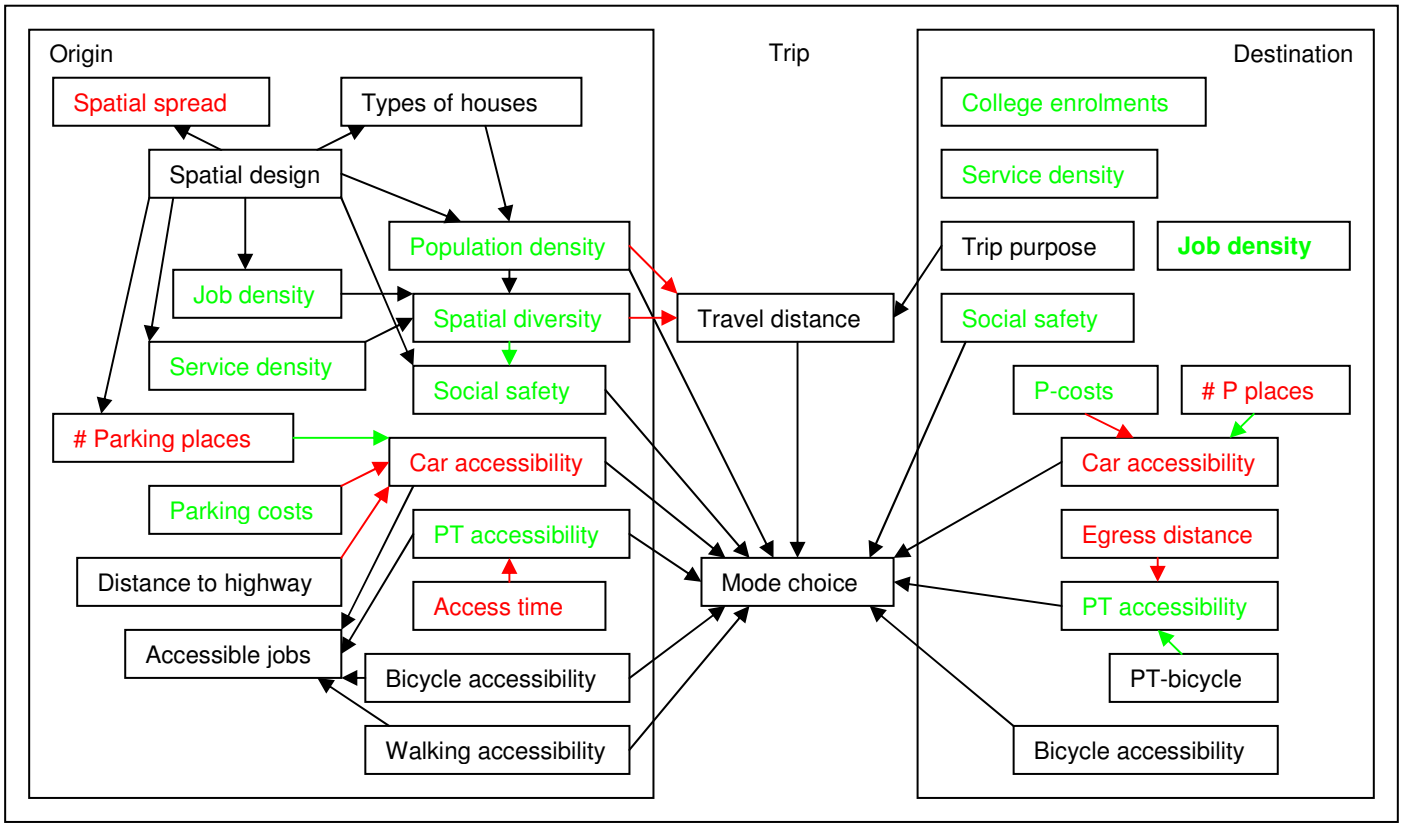


Figure A.2 Built environment variables related to the demand of bus transport

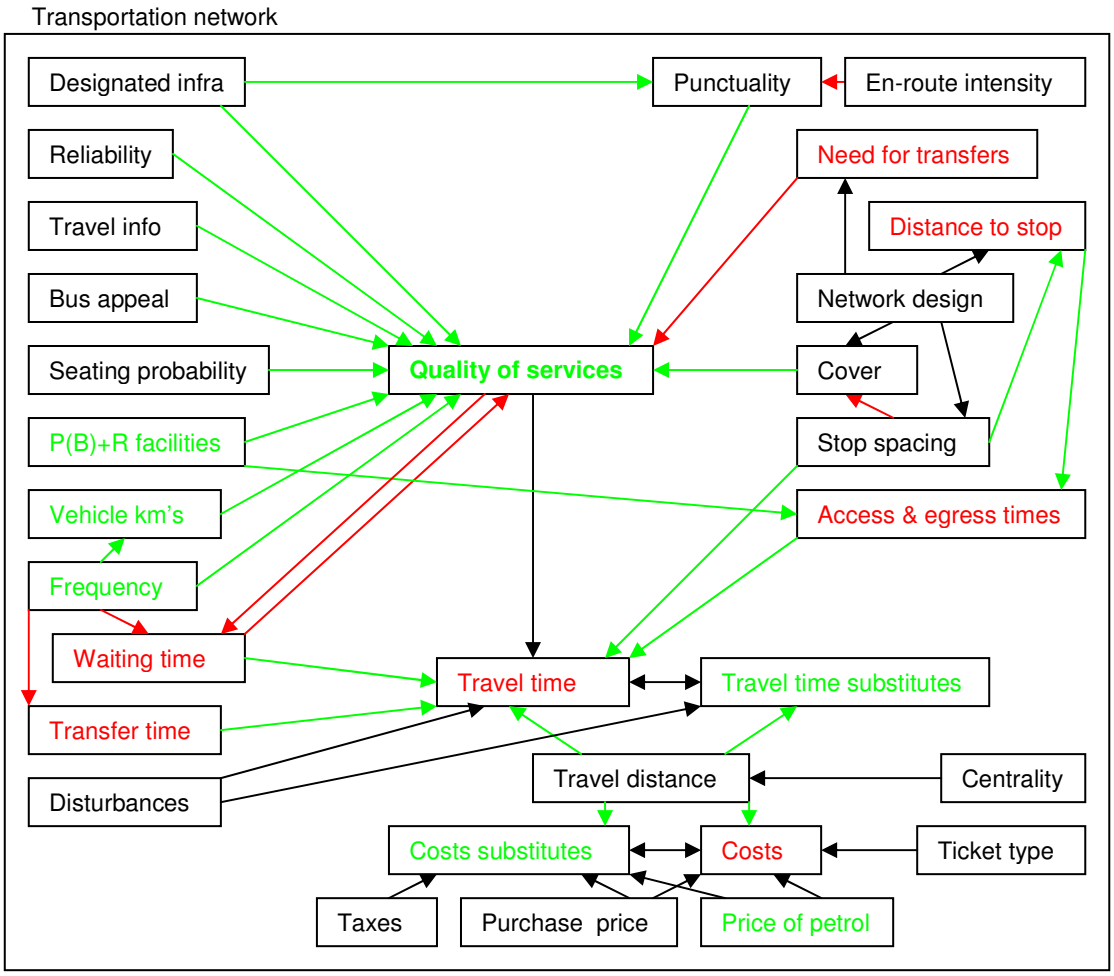


Figure A.3 Transportation network variables related to the demand of bus transport

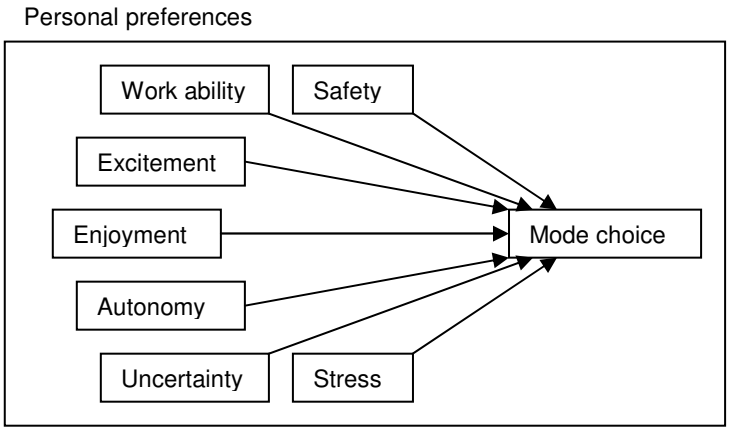


Figure A.4 Personal preference variables related to the demand of bus transport

Explanation of the figures:
The figures can best be explained using the example at the right:
The variables that are thought to have a positive effect (positive elasticity) on the demand for bus transport are written in green (like the percentage of students). Variables that are thought to have a negative elasticity are written in red (like income and car ownership). Variables that are very important according to research are written bold (such as car ownership).

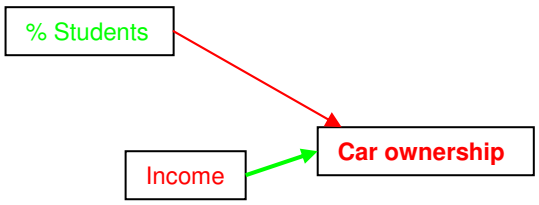


Figure A.5 Example

When a relation between two variables is positive (an increase in the value of one results in an increase in the value of the other) a green arrow is placed between the two variables (an increase in income is hence thought to result to an increase in car ownership). For a negative relation between two variables a red arrow is used (an increase in the percentage of students results in reduced car ownership in a certain area). Relations that are stronger according to literature have thicker arrows than others (the relation between income and car ownership is very important according to literature). When variables are written in black the relation with the demand for bus transport is not present, the relation is not straightforward or the relation is not known. When an arrow is black the relation between the variables cannot be described as: the increase/decrease of the one variable, leads to a increase/decrease of the other variable.

Appendix B: Subdivision into zones

Zone name	Neighbourhoods	Lines
City	City	1, 2, 3, 4, 5, 6, 7, 17, 18, 20, 60, 61, 62, 64, 73, 74, 76
	Getfert	
	Boddenkamp	
Hogeland	Hogeland-Zuid	5
	Varvik-Diekman	
Boswinkel	Cromhoffsbleek-Kotman	1
	Boswinkel-De Braker	
Stadsveld 4	Horstlanden-Stadsweide	4
	Stevenfenne	
	't Zwering	
Stadsveld 5	Veldkamp-Getfert-West	5
	Pathmos	
	Stadsveld-Zuid	
	Elferink-Heuwkamp	
	Stadsveld-Noord-Bruggert	
	't Zwering	
Twekkelerveld	Tubantia-Toekomst	1
	Twekkelerveld	
Enschede Noord	Lasonder, Zeggelt	2
	Bolhaar	
	Roombeek-Roomveldje	
	Mekkelholt	
	Deppenbroek	
	Voortman-Amelink	
UT	Drienerveld-U.T.	1
Stokhorst	De Laares	6
	Schreurserve	
	Ribbelt-Ribbelerbrink	
	Park Stokhorst	
	Stokhorst	
Stroinkslanden	Stroinkslanden Noord-Oost	4
	Stroinkslanden-Zuid	
	Stroinkslanden Noord-West	
Wesselerbrink 1	Wesselerbrink Noord-Oost	1
	Wesselerbrink Zuid-Oost	
	Wesselerbrink Zuid-West	
	Wesselerbrink Noord-West	
Wesselerbrink 2	Wesselerbrink Zuid-Oost	2
	Wesselerbrink Zuid-West	
Helmerhoek	Helmerhoek-Noord	2
	Helmerhoek-Zuid	
Bedrijven West	Industrie- en havengebied	6
	Marssteden	
	Koekoeksbeekhoek	
Glanerbrug	Glanerveld	3
	Bentveld-Bultserve	
	Schipholt-Glanermaten	
	Eekmaat	
	Oikos	

Glanerbrug West	Sleutelkamp	3
	't Weldink	
	De Leuriks	
	De Slank	
	Dolphia	
	Eekmaat West	
Boekelo	Dorp Boekelo	6
Haaksbergen	Haaksbergen Kern-1	20, 73, 74, 76
	Haaksbergen Kern-2	
	Haaksbergen Kern-3	
	Haaksbergen Kern-4	
	Veldmaat 1	
	Leemdijk	
	Industriegebied West	
	Industriegebied Brammelo	
	Haaksbergen Kern-Centrum	
Losser	Losser-Oost	61
	Losser-West	
	Bedrijfsterrein Zoeker Esch- De Pol	
Overdinkel	Overdinkel kern	61
Haaksbergen-2	Haaksbergen Kern-1	53, 59
	Haaksbergen Kern-2	
	Haaksbergen Kern-3	
	Haaksbergen Kern-4	
	Veldmaat 2	
	Zienesch	
	de Pas	
	de Els	
	Wolferink 1 en 4	
	Wolferink 2	
	Wolferink 5	
	Wolferink 3	
	Hassinkbrink	
	Industriegebied West	
	Haaksbergen Kern-Centrum	
Losser-2	Losser-Oost	64
	Losser-West	
	Bedrijfsterrein Zoeker Esch- De Pol	
Overdinkel-2	Overdinkel kern	64

Appendix C: Relative accessibility

The relative accessibility of the different modes from a certain zone is one of the variables included in the research. In this appendix it is described how this variable is determined.

How many jobs/inhabitants are accessible per mode?

Accessibility in this context is ”the ability to reach activities or locations by means of a travel mode” (Geurs & van Wee, 2004). For this research the number of jobs and inhabitants that can be accessed by the different modes are used to compare the different zones with each other. With the Regional traffic model it is possible to determine those numbers of people and jobs that can be accessed or reached within a certain amount of time. For the comparison it is important what reasonable travel times are for the different modes. Another question is whether jobs/inhabitants located on different distances from an origin should be weighed equally. The designer of the Regional Traffic Model (Goudappel Coffeng) uses experience based curves to weight jobs with certain travel times towards the origin. These curves are mode-specific, which means that they are different for all modes. Because Goudappel Coffeng obtained this information because of their long-lasting experience with transport modelling they do not want to publicise their exact numbers, but the global (relative) curves are shown in figure C.1. In this figure it can be seen people are more likely to travel longer by Public Transport than by car and longer by car than by bicycle. The number of accessible jobs/inhabitants is a summation of the number of jobs/inhabitants in an accessible zone multiplied by the worth of the travel time for a trip form the origin zone to the accessible zone.

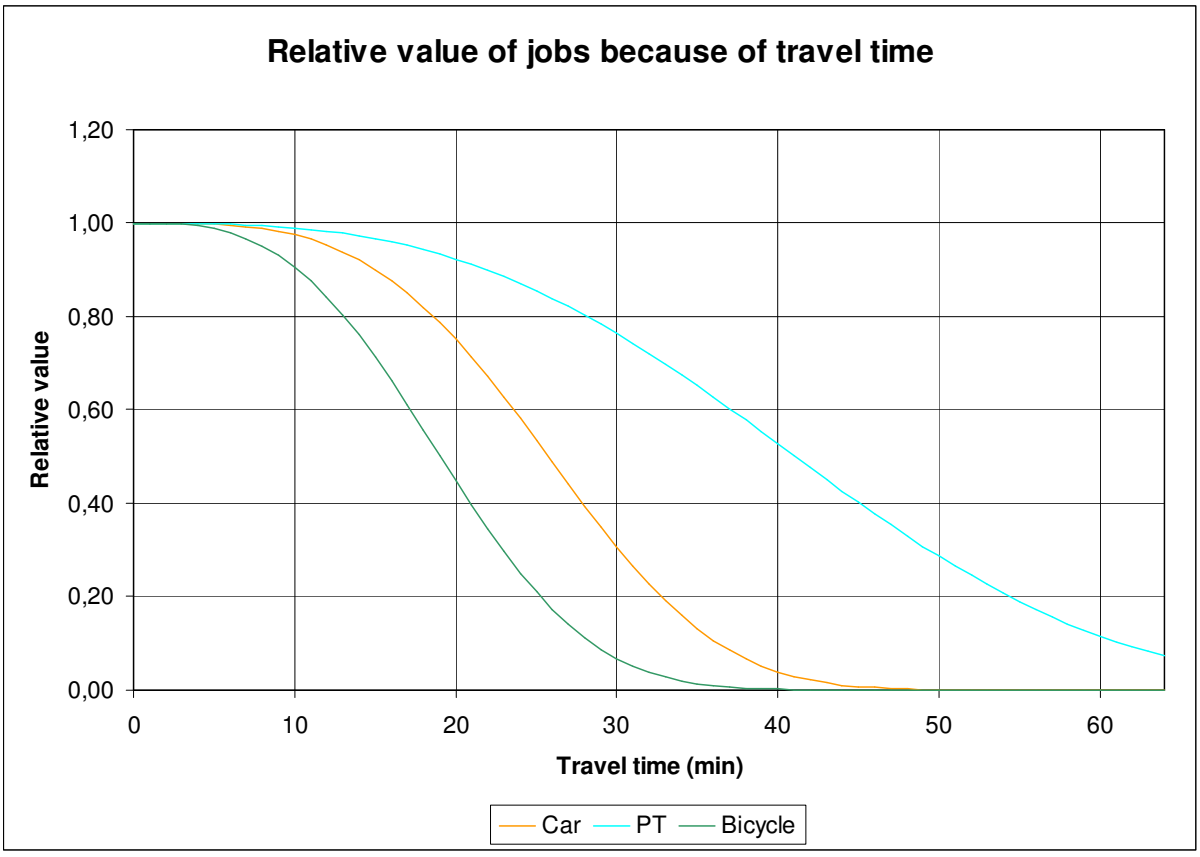


Figure C.1 Travel time-relative value curves

Travel times of public transport and bicycle are more or less constant over time. Travel times by bicycle are not influenced by increased bicycle intensities whereas bus transport can be prioritised at intersections to keep delay due to congestion to a minimum. Therefore the travel times for these modes that are used to determine the number of accessible jobs and inhabitants are whole day travel times. Car travel times are higher in peak periods because then more cars use the road. This higher intensity makes that the travel time on the links becomes longer. Because many trips form the neighbourhoods are performed in the morning peak hour the morning peak hour travel times in the

model are also used to determine the number of jobs and inhabitants that can be accessed by car travel.

Why Public Transport instead of Bus?

The research only considers the demand for bus transport. The Regional Traffic Model does not differentiate between bus and train trips. People either travel by Public Transport or by another mode. It is hence impossible to determine the bus accessibility alone. Therefore the model can only determine the accessibility of Public Transport as a whole. Regional and urban bus transport and train transport however are related because bus transport is one of the feeders of (regional and interregional) train.

Implications

Using the accessibility by Public Transport as a whole does have some implications and disadvantages.
Most important disadvantage is that because large distances in relatively short time can be travelled by train, the train probably has most influence in the number of jobs or inhabitants accessible. This leads to high numbers of accessible jobs/inhabitants by PT in city centres and relatively an overestimation of the accessibility of jobs/inhabitants by bus transport in these areas.
Including train travel in the PT accessibility also means that another curve is used for the value of jobs that can be accessed within a certain travel time than when only bus travel would be used. The reason behind this difference is that train travel is better suited for longer distance travel than bus travel, because travelling by bus is less comfortable and slower. The curve that represents the value of a job/inhabitant given a certain travel time (figure C.1) of bus transport therefore can also be more towards the car curve, which would indicate that the number of jobs that can be accessed is overestimated. However this overestimation takes place for all zones which might reduce the negative effects of it.

Other ways to determine the accessibility by bus transport however are not at hand. Because no better information on the bus accessibility is present this measure is used. It however may turn out that the related variable is unusual as a determinant for the demand of bus transport.

The relative accessibility

The absolute accessibility is interesting for location choice behaviour, but not for modal choice behaviour. For modal choice it is interesting how many jobs and inhabitants are accessible by public transport in relation to the ones that are accessible by car and the ones that are accessible by bicycle. The relative accessibility of the bus is defined as follows:

$$\text{Relative car job accessibility} = \frac{\#\ jobsCar}{\#\ jobsPT} \qquad \text{(C.1)}$$

For the relative accessibility by bicycle the number of jobs/inhabitants by car is replaced by the number by bicycle.

Model specifications

The Regional model has some specifications that has implications for the reliability and correctness of the data.
The zones are for instance connected to the network using connector links. All inhabitants and links are therefore modelled to be on the point where the zone is connected to the model. In reality this is off course not the case. The travel time of the different modes therefore is not exact.
For Public Transport the previously mentioned fact also has other influence on the accessibility. The reason for this is that for this mode access & egress transport is needed. The location of the connector therefore determines the access and egress times, and consequently the number of jobs and inhabitants that can be accessed within certain time, to a large extend. The mode of access transport is determined by the distance to the nearest stop. For short distances walking is used, for longer distances bicycling and for long distances the car is used. When certain zones have connectors very close to a bus stop and others not the relative accessibility of the bus is quite different in these zones. In reality it can however be that people can access bus stops very easy (because of footpaths or small alleys) in the zones that are connected to the network far off the bus stop. This influences the reliability of the data negatively.

Outcomes

The outcomes of the previous steps are now described.

Accessibility per mode

First an overview is given of the accessibility of the different modes in Twente. In this way it can be checked whether the model gives reasonable results. In figure C.2.a-c the job accessibility of the three modes is given. For car (C.2.a) it can be seen that zones near highways and provincial roads have the most jobs accessible. This is what one would expect because large distances can be travelled in a short time over these roads. For bicycle (C.2.b) the zones that have most jobs accessible are located in or near urban areas. The reason for this is that people do not travel very far by bicycle. Within urban areas the concentration of jobs is usually highest which means that most jobs can be accessed by bicycle from urban zones. The further to the city centre the zone is, the more urban area (and jobs) can be accessed within reasonable time. Between Enschede and Hengelo also a high number of jobs can be accessed because both cities can be reached within reasonable time and because the area itself has a high concentration of jobs. For Public Transport (C.2.c) most jobs can be accessed from zones that are near train stations. Also from zones that are near (important) bus lines (Haaksbergen, Losser) more jobs can be accessed than from zones that are far away from bus stops and lines. The outcomes are all quite reasonable.

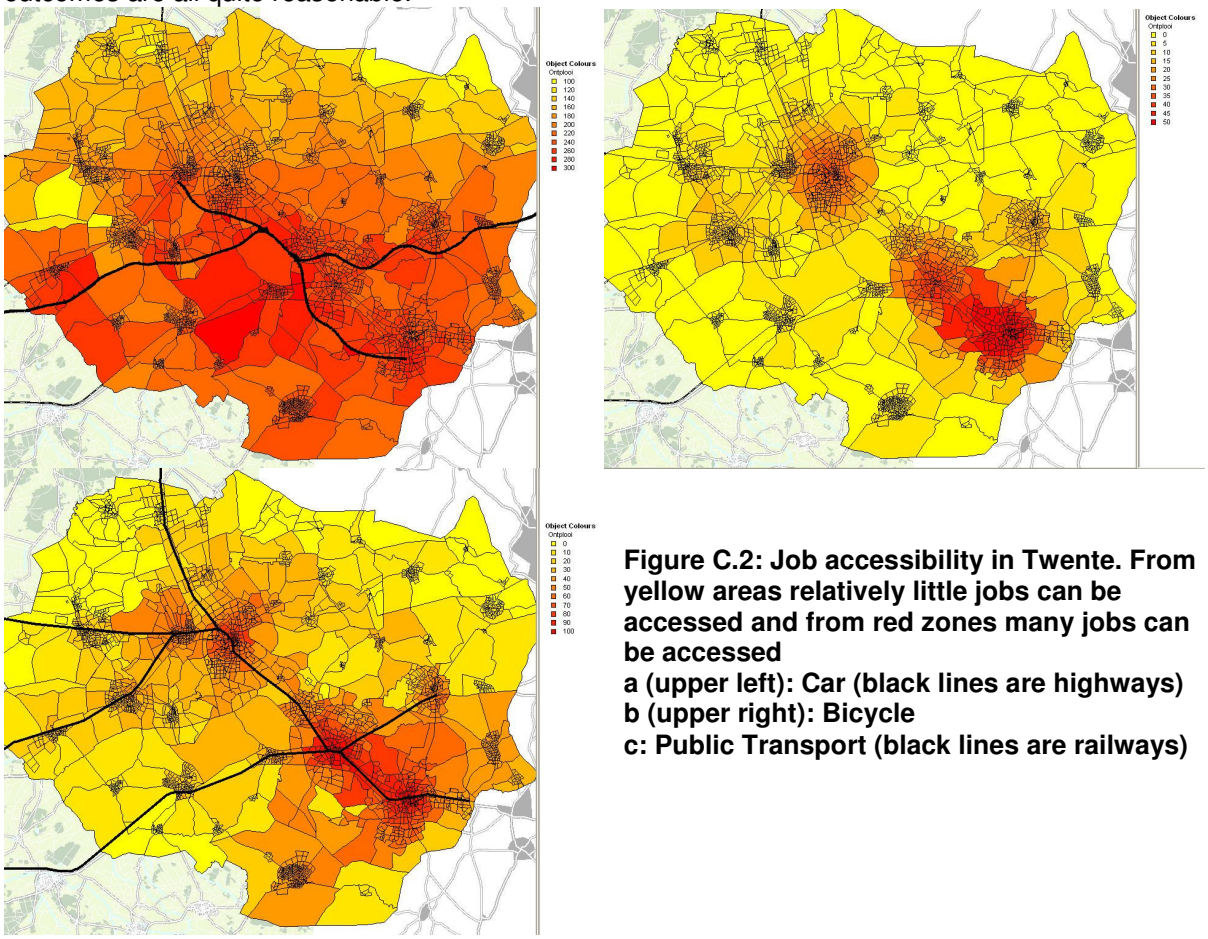


Figure C.2: Job accessibility in Twente. From yellow areas relatively little jobs can be accessed and from red zones many jobs can be accessed
a (upper left): Car (black lines are highways)
b (upper right): Bicycle
c: Public Transport (black lines are railways)

In figure C.3 the accessibility of jobs by car and public transport are plotted together. In the figure a subdivision is made for the different zones to the municipality they are located in. It can be seen that in Enschede the accessibility is quite spread. In the other municipalities the values are located closer to each other, which means that the values of both variables are comparable. This difference is pretty much related to the size of the municipalities. While Enschede is bigger, more lines are exploited and a train station is present in the centre the opportunities for using public transport are very different for the different zones in the municipality. On the other hand, in Losser only two lines are exploited and only two regional roads towards Enschede and Oldenzaal are present. Therefore the only difference in the accessibility is the travel time in Losser itself to reach the bus or the regional roads. The number of accessible jobs from the zones therefore is close to each other. Because both Losser and Overdinkel

are taken into account two groups of dots can be seen in figure C.3. In the municipality of Haaksbergen only the village of Haaksbergen is included which is the explanation fo the fact that there is only one group of dots.

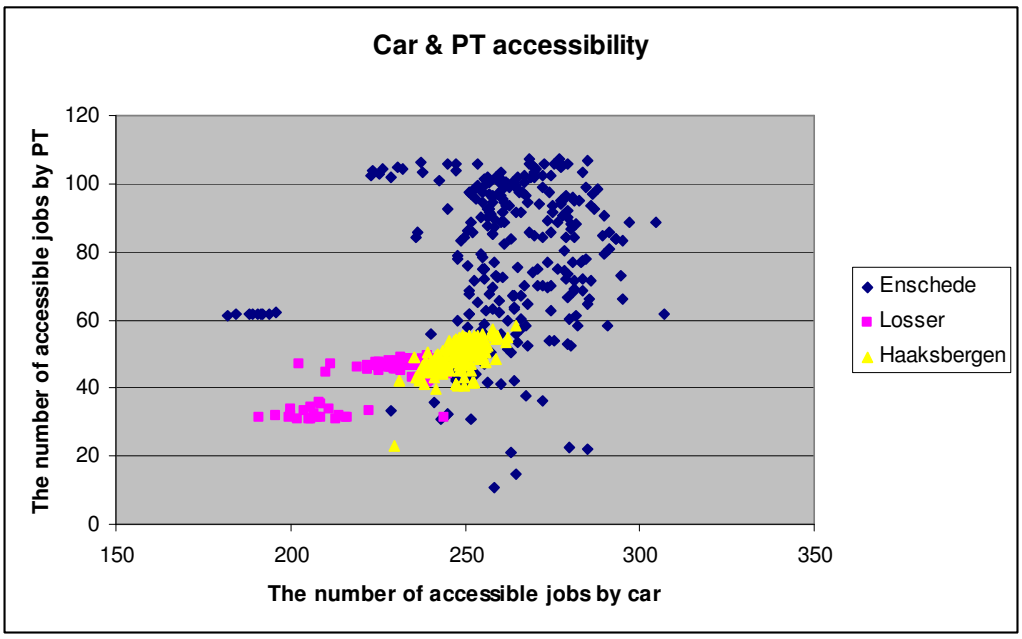


Figure C.3 Municipality specific job accessibility by car and public transport

Relative accessibility

For all the zones in the study area the accessibility of Public Transport relative to car and bicycle is determined. On average most jobs and inhabitants can be accessed by car. A reason for this is that travel by car is relatively fast and that relatively much infrastructure is present. The second most jobs and inhabitants can be accessed by Public Transport because it (especially train) is faster than bicycling and because the reasonable travel time by PT is longer than by bicycle.
The relative accessibility of the different modes varies over the different zones. From some zones about the same amount of jobs and inhabitants can be accessed by public transport as by car while from other zones nearly no jobs/people can be accessed by public transport. Reasons for this are that some zones are near train stations (many jobs/inhabitants accessible by PT) while others are far away from a bus line (few jobs/inhabitants accessible by PT). In figure C.4a it can be seen how the relative job accessibility varies over the zones. The zones that are yellow have relatively large numbers of jobs accessible by PT while pink zones can access a lot of jobs more by car than by Public Transport. In red areas the number of jobs accessible by car exceeds the number of jobs accessible by PT. This can either be because PT service is very poor in the zone or because car opportunities are very high. It can be seen that areas around the railway stations (railways are the black lines) have lower relative accessibility of cars and areas around highways have relatively high numbers of jobs available.

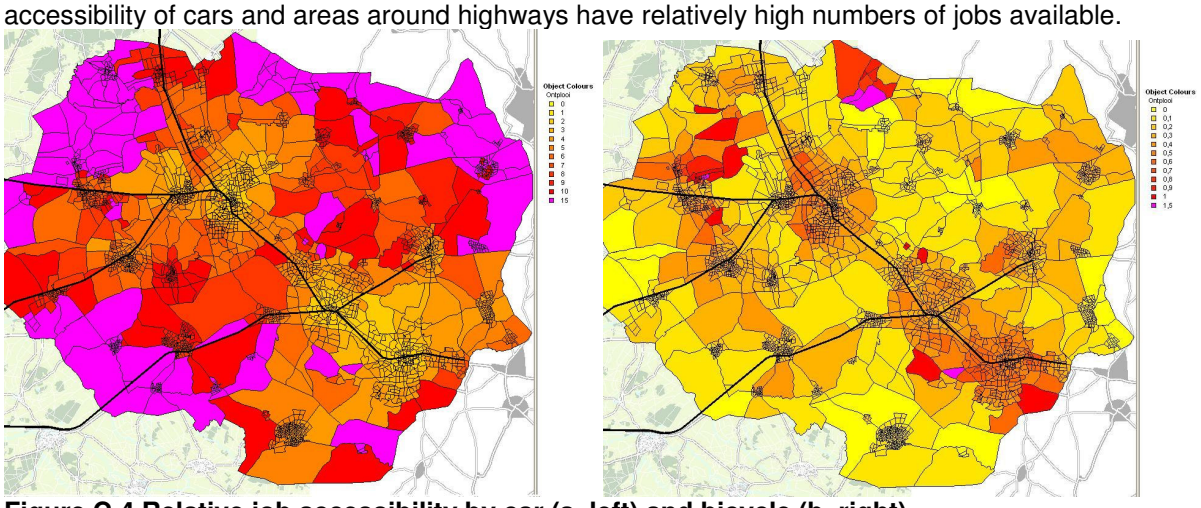


Figure C.4 Relative job accessibility by car (a, left) and bicycle (b, right)

The relative number of jobs accessible by bicycle compared to Public Transport is shown in figure C.4b. In this figure it can be seen that, although areas in urban areas have relatively good access to Public Transport, their relative job accessibility by bicycle is higher than in rural areas. The areas that are very red or purple in this case are the zones that have very poor PT-access.

Correlation

The outcomes presented all considered the area of Regio Twente as a whole. Now the focus goes back to the research area (the municipalities of Enschede, Haaksbergen and Losser). For these areas it is important to check whether there is correlation between the relative accessibility of jobs and inhabitants. First it however is important to determine whether or not there are outlier values in the data.

Outliers

In all datasets it is possible that there are outliers. Outliers are values that highly deviate from the sample mean. When these values remain in the sample they can highly influence the reliability of the relationships that are concluded from the data. The outliers are identified using Chauvenet's criterion (Taylor, 1997)[21]. In table C.1 an overview of the number of outliers for all the four categories of relative accessibility is given.

	Car/PT	Bicycle/PT
Jobs	7	4
Inhabitants	6	4

Table C.1 Number of outliers

Analysis

A reason for the extreme values can be that some areas are very bad accessible by Public Transport. In this way the denominator of the relative accessibility becomes very small which leads to a high relative accessibility. In figure C.4 some of the outliers are the pink area West of Enschede and the red area West and South-East of Enschede. The pink area (Marssteden business park) however has a bus line running through it. The low number of jobs accessible by Public Transport therefore seems to be a model error. The other two areas do not have PT lines and, because they are close to Enschede, it is plausible that the relative accessibility by bicycle is high.

Correlation

For the research it is interesting to look at potential correlation between the accessibility of jobs and inhabitants. Because both jobs as inhabitants are concentrated in urban areas it is plausible that this correlation exists. When the two variables are highly correlated they should not both be included in the model. For both car and bicycle it is checked whether there is correlation. The outliers are excluded from the data to exclude model errors.
In figure C.5 the relative car accessibility of jobs is plotted against the relative car accessibility of inhabitants. It can be seen that there is a clear linear relation between the two variables. It can be seen that the residual sum of squares is only about 2%. This means that the model fits nearly perfect. Another measure for correlation is the Pearson product-moment correlation coefficient (PMCC) (see chapter 3.2.2 for a more detailed information about correlation). The value if this coefficient is 0,983 which means that the correlation is nearly one (which would mean perfect correlation). Kuby et al. (2004) describe that the danger level of a correlation coefficient is 0.7. When variables have a Pearson correlation coefficient that is higher than this danger value one of the two variables should be excluded form the regression analysis. Therefore one of the variables should be left out of the research. Because most trips are trips to work, school or shops, which all are related to jobs in the destination area, it is chosen to keep the relative accessibility of jobs in the research and to exclude the relative accessibility of inhabitants.

[21] The Chauvenet's criterion states that values are outliers when the probability of obtaining the particular deviation from the mean is less than $1/(2n)$.

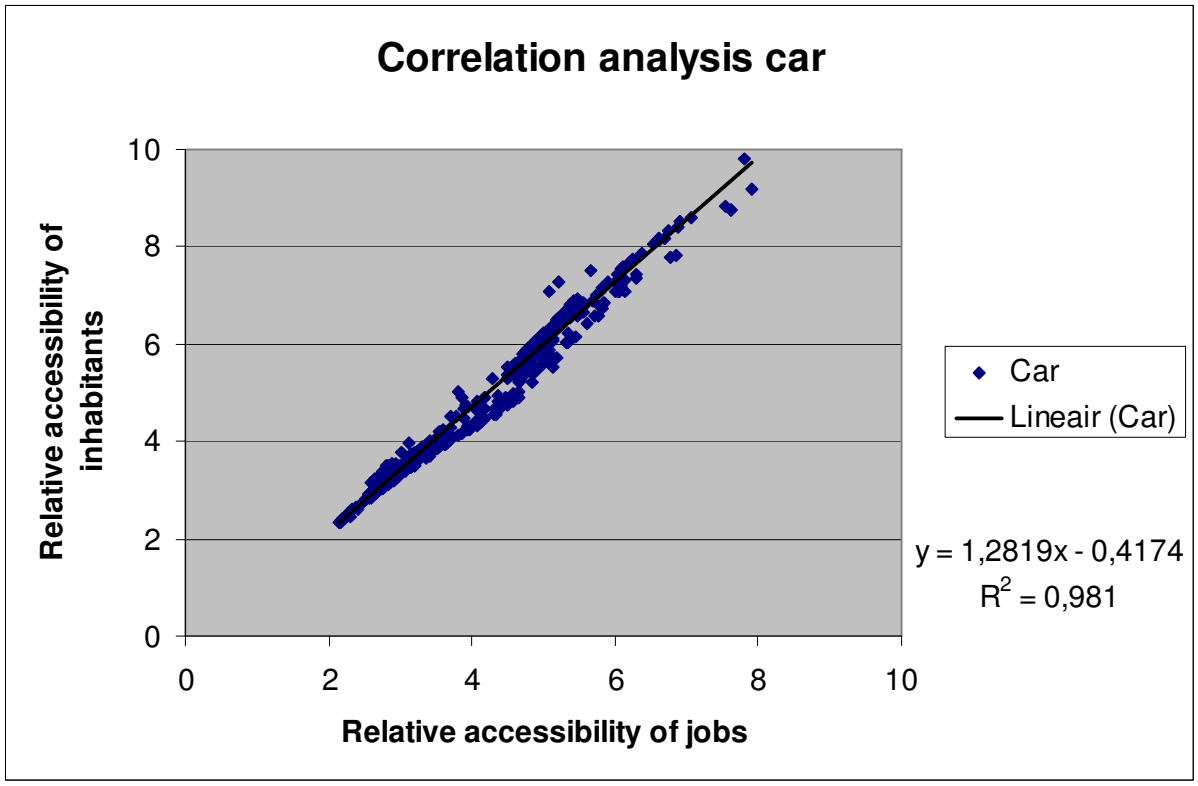


Figure C.5 Correlation analysis car

In figure C.6 the relative bicycle accessibility of jobs and inhabitants are plotted against each other. Again it can be seen that there is a high correlation between the two variables. The R^2 is 0.9202 in this case and the PMCC is even 0.957. Again these values are high above the danger level. The relative inhabitant accessibility by bicycle therefore is also not taken into account in the research.

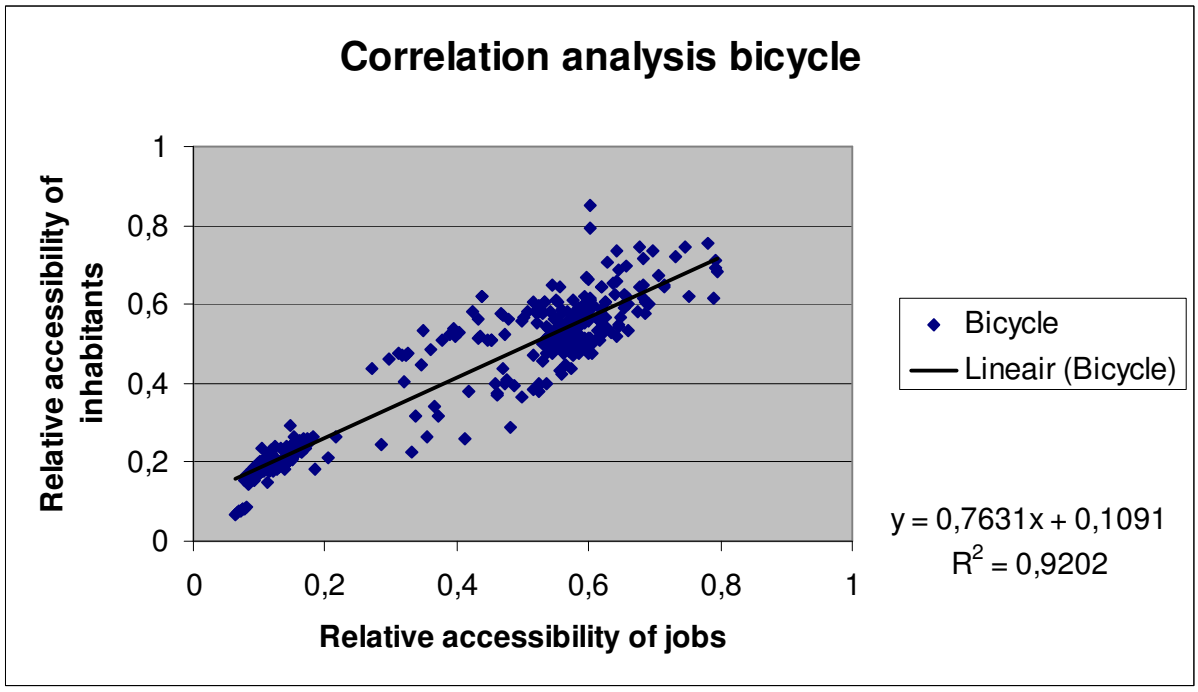


Figure C.6 Correlation analysis bicycle

Appendix D: Distance to stop

In this appendix the determination of the average distance to a stop is described. It is determined with the ArcGIS Software available at the ITC in Enschede.

Base Data

Earlier it was mentioned that the (boarding) data that are available do not allow to take all stops as separate research subjects. The consequence of this fact is that not the number of inhabitants & jobs in the catchment area of a stop is interesting, but the number of inhabitants & jobs in a zone and the average distance to a stop in that zone.
In order to determine the average distance to a stop it is off course important to have the stops and the different zones in a digital format. The stops come from the Regional Traffic Model Twente (RVM) and the zones come from the neighbourhood map of Dutch Statistics.

When the stops are imported into the ArcGIS software some problems can be seen. The stops are those of before the line alterations of line 2 towards Enschede North in 2008. Because of that some stops along the new designated bus lane are not included in the data. Also in Glanerbrug and Overdinkel some stops are missed and others are included while they are not in service anymore. There are also some stops that are wrongly located or located on two locations because the two (or more) stops on both sides of the roads are far from each other.
Fortunately there is another source with the exact location of all bus stops in the region that are serviced by Connexxion. In this document the X and Y (RD-)coordinates[22] of all stops (on all sides of the road) are stored. With the previously mentioned calculation method (paragraph 7.6.4) the location of the representative of all stops are determined. For the previously mentioned stops these coordinates are used to give the correct graphical representation in ArcGIS.

The zones of DS are not identical to the zones of the number of boardings. The neighbourhoods are smaller and the districts are larger than these zones. To make aggregation possible the average distance to a stop in the neighbourhoods are determined.

Distance to a stop

The shortest distance towards multiple features (points, lines, polygons) in ArcGIS can be determined with the function Euclidian Distance. This function determines for all 10x10 metre areas in the research area the straight line distance towards the nearest bus stop.

Or in a formula: $DtS_i = \sqrt{(x_i - x_s)^2 + (y_i - y_s)^2}$

Where:

DtS_i = Euclidian distance to stop of location i

x_i = x-coordinate of location i

x_s = x-coordinate of the nearest stop

y_i = y-coordinate of location i

y_s = y-coordinate of the nearest stop

The outcome is a raster (pixel) layer with a value for the distance towards a stop for the whole area a pixel represents. In figure D.1 the result of this step is visualised. Also the stops themselves and the neighbourhood borders are shown on this figure. It can be seen that around stops circular areas are coloured equally which means that they are equally far from the stop. The more dark blue an area is the farther it is located away from the nearest stop.

[22] RD-coordinates: RijksDriehoek coordinates; Coordinates that are used in The Netherlands as the basis for the geographical representation.

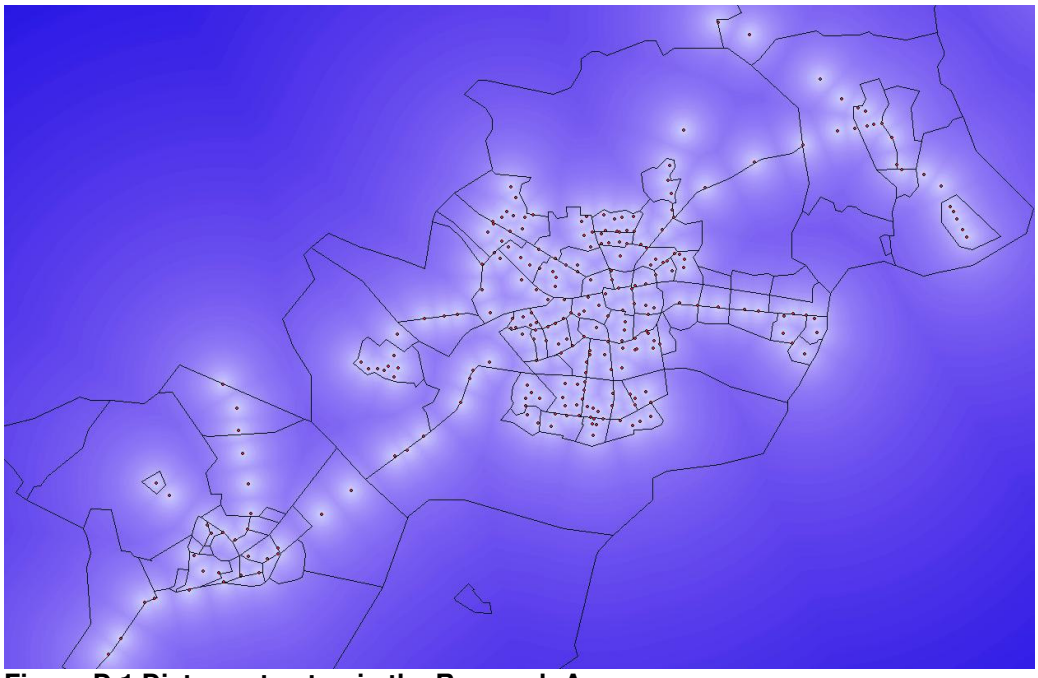


Figure D.1 Distance to stop in the Research Area

Average distance to stop per neighbourhood

Now that for all 10x10 areas in the research area the distance towards a bus stop is known the average for the different neighbourhoods can be determined. This can be done with the Zonal Statistics function in ArcGIS. This function determines for all neighbourhoods the minimum, maximum and mean of a raster layer alongside with the spread and the sum of all raster areas in the zone. The mean (average) value of the distance towards a stop for the different zones is shown in figure D.2.

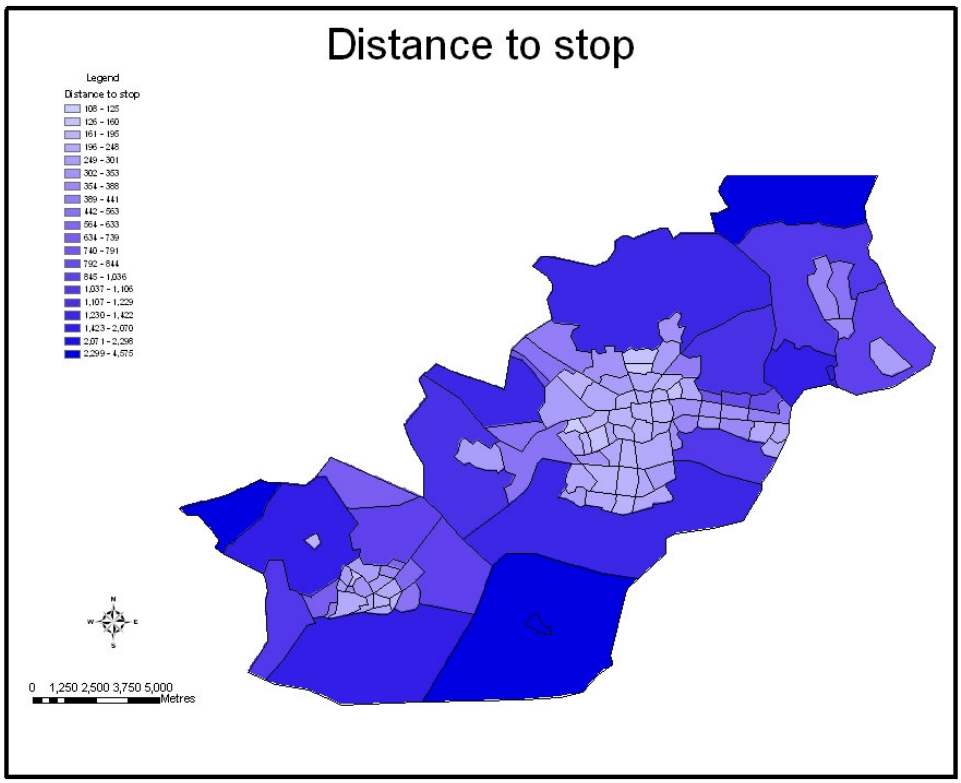


Figure D.2 Distance to stop on a neighbourhood level

In figure D.2 the neighbourhoods that are the lightest blue are the zones that have the smallest average distance towards a bus stop. It can be seen that some neighbourhoods in Enschede have tha smallest average distance towards a bus stop. The other neighbourhoods in Enschede can be compared to neighbourhoods in Haaksbergen as far as the average distance towards a bus stop is concerned and the neighbourhoods in Losser are clearly darker which means that the distance towards a stop is higher there. The areas around Buurse (the Western part of Haaksbergen) and North of Losser have the largest distances towards a (normal) bus stop.

Average distance to stop per zone

With the values of the neighbourhoods the values of the aggregated zones can also be calculated. This is done by weighing the different neighbourhoods based on their area. A neighbourhood that is twice as large as another one therefore counts double in determining the average distance towards a stop in the zone.

Appendix E: Correlation in urban and rural areas

In table E.1 the correlation between the different independent variables and the dependent variable (the number of boardings per inhabitant) in a zone is showed. Three zones are not taken into account: the city centre, the university and the harbour.

Variable	Overall	Urban	Rural
Car ownership	0.026	0.012	0.375
Business car ownership	0.148	0.112	0.505
Income	0.265	0.534	-0.265
% of students	0.754	0.740	0.194
% of people aged 15-24	0.804	0.807	-0.111
Inhabitants	-0.361	-0.473	-0.156
Jobs / Inhabitant	0.187	0.149	0.295
College enrolments / Inh.	0.829	0.849	-
Relative car accessibility	-0.310	-0.159	0.058
Relative bicycle accessibility	0.172	-0.255	-0.219
Travel time to city centre	-0.046	0.132	0.667
Relative Travel Time car	-0.029	0.431	-0.265
Relative Travel Time bicycle	-0.175	0.420	-0.156
Population density	-0.283	-0.460	-0.126
Job density	0.244	0.157	0.189
Address density (OAD)	-0.163	-0.530	0.012
Frequency	0.406	0.248	0.657
Punctuality	0.116	-0.152	0.456
Train Station	-	-	-
Distance to stop	0.226	0.461	-0.191

Table E.1 Correlation coefficients with the dependent variable, for urban and rural zones

Appendix F: Correlation between independent variables in Neighbourhoods

	Car ownership	Company car	Income	Student	People aged 15-24	Inhabitants	Jobs	College enrolments	Access car Jobs	Access cycle Jobs	Frequency	Punctuality	Distance to stop	Travel Time centre	Rel TT Car	Rel TT Bicycle	Address density	Population density	Job density
Car ownership	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Company car	0.491[23]	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Income	-0.248[23]	-0.394[23]	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Student	-0.116[23]	-0.056[23]	0.094[23]	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
People aged 15-24	-0.227	-0.142	-0.141	0.676	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Inhabitants	-0.296	-0.290	-0.096	0.003	0.010	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Jobs	0.311	0.314	-0.096	0.077	0.088	0.005	x	x	x	x	x	x	x	x	x	x	x	x	x
College enrolments	-0.167	-0.039	-0.096	0.841	[24]	0.309	0.961	x	x	x	x	x	x	x	x	x	x	x	x
Access car Jobs	0.157	0.098	0.042	-0.613	-0.456	-0.165	-0.356	-0.113	x	x	x	x	x	x	x	x	x	x	x
Access cycle Jobs	-0.103	-0.041	-0.127	0.372	0.349	0.138	0.186	-0.040	-0.624	x	x	x	x	x	x	x	x	x	x
Frequency	-0.092	-0.073	-0.024	0.278	0.302	0.042	0.606	0.507	-0.310	0.188	x	x	x	x	x	x	x	x	x
Punctuality	-0.248	-0.224	-0.056	0.113	0.108	0.040	-0.191	-0.394	-0.242	0.573	-0.050	x	x	x	x	x	x	x	x
Distance to stop	0.165	0.179	0.197	-0.172	-0.110	-0.257	-0.077	0.165	0.316	-0.237	-0.159	-0.155	x	x	x	x	x	x	x
Travel Time centre	0.123	0.056	0.120	-0.436	-0.406	-0.272	-0.443	-0.098	0.719	-0.680	-0.485	-0.176	0.142	x	x	x	x	x	x
Rel TT Car	0.115	0.093	0.113	-0.349	-0.265	-0.135	-0.059	0.429	0.562	-0.801	-0.047	-0.554	0.297	0.393	x	x	x	x	x
Rel TT Bicycle	0.136	0.073	0.146	-0.360	-0.319	-0.179	-0.079	0.233	0.597	-0.944	-0.055	-0.661	0.244	0.556	0.907	x	x	x	x
Address density	-0.337	-0.305	-0.275	0.397	0.330	0.282	0.275	0.143	-0.735	0.535	0.421	0.359	-0.546	-0.626	-0.553	-0.541	x	x	x
Population density	-0.406	-0.396	-0.367	-0.013	0.069	0.515	-0.179	0.152	-0.236	0.163	0.078	0.304	-0.587	-0.143	-0.191	-0.202	-0.350	x	x
Job density	0.099	0.111	-0.083	0.319	0.348	-0.066	0.865	[24]	-0.345	0.127	0.744	-0.125	-0.167	-0.420	-0.061	-0.034	-0.046	-0.098	x

Table F.1 Correlation coefficients of relations between different independent variables

[23] Determined for all neighbourhoods in the study area rather than only neighbourhoods with a certain number of boardings
[24] Cannot be determined because of the low number of zones with college enrolments

Appendix G: Correlation between variables in zones (research objects)

Correlations																			
	Car ownership	Business car ownership	Income	% of students	% of people aged 15-24	Inhabitants	Jobs per inhabitant	College enrollments per inhabitant	Rel access car	Rel access bicycle	Frequency	Punctuality	Distance to stop	Travel time by bus	Relative travel time car	Relative travel time bicycle	Address density	Population density	Job density
Car ownership	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Business car ownership	.667**	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Income	.805**	0.359	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
% of students	-.561*	-0.321	-0.334	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
% of people aged 15-24	-.726**	-.564**	-.624**	.797**	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Inhabitants	-0.315	-0.074	-0.243	0.222	0.135	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Jobs per inhabitant	0.103	.524*	0.049	-0.021	-0.194	-0.294	x	x	x	x	x	x	x	x	x	x	x	x	x
College enrollments per inh.	-0.025	-0.108	0.138	0.047	0.021	0.056	-0.098	x	x	x	x	x	x	x	x	x	x	x	x
Rel access car	.699**	0.327	.471*	-.750**	-.691**	-0.283	-0.160	-0.236	x	x	x	x	x	x	x	x	x	x	x
Rel access bicycle	-.722**	-.494*	-0.369	.601**	.563**	0.010	0.120	0.173	-.720**	x	x	x	x	x	x	x	x	x	x
Frequency	-.651**	-0.167	-.570**	0.166	0.247	0.296	0.125	-0.232	-0.324	0.365	x	x	x	x	x	x	x	x	x
Punctuality	-0.386	-0.314	-0.216	0.328	0.374	-0.025	-0.155	0.288	-0.434	.505*	0.180	x	x	x	x	x	x	x	x
Distance to stop	.586**	0.373	0.419	-.612**	-.738**	-0.058	0.226	-0.043	.599**	-.588**	-0.171	-.448*	x	x	x	x	x	x	x
Travel time by bus	.589**	.475*	0.352	-.613**	-.587**	-0.083	-0.204	-0.164	.784**	-.709**	-0.170	-0.111	0.277	x	x	x	x	x	x
Relative travel time car	.626**	.465*	0.332	-.507*	-.503*	0.085	0.014	-0.341	.582**	-.819**	-0.197	-.530*	.634**	0.430	x	x	x	x	x
Relative travel time bicycle	.740**	.553*	0.399	-.535*	-.534*	0.014	-0.007	-0.265	.670**	-.942**	-0.371	-.577**	.604**	.591**	.930**	x	x	x	x
Address density	-.790**	-0.394	-.662**	.615**	.706**	0.369	0.029	0.178	-.842**	.723**	0.316	.495*	-.645**	-.619**	-.717**	-.745**	x	x	x
Population density	-.775**	-.571**	-.698**	0.387	.692**	.492*	-.491*	0.053	-.534*	.458*	.479*	.464*	-.603**	-0.308	-0.370	-.459*	.681**	x	x
Job density	-0.370	0.108	-0.397	0.285	0.415	-0.040	.707**	-0.045	-.524*	0.313	0.278	0.098	-0.253	-0.427	-0.218	-0.239	.519*	0.104	x
**. Correlation is significant at the 0.01 level (2-tailed).																			
*. Correlation is significant at the 0.05 level (2-tailed).																			

Table G.1 Correlation coefficients of relations between different independent variables

Appendix H: Regression

In this appendix the overview of the different regression steps is given.

In the table below the different colours mean the following:
In column "Adjusted R^2": Green: highest value so far. Red: lower value than currently highest (with high significance).
In all other columns: Green: Standard error smaller than 2% two-tailed (1% one-tailed). Yellow: Between 2% and 10% (1% and 5%). Orange: between 10% and 20% (5% and 10%). Red: Higher than 20% (10%).

Model	Adjusted R^2		Constant	% of people aged 15-24	% of students	Business car ownership	Car ownership	Distance to stop	Frequency	Income	Jobs per inhabitant	Punctuality	Rel access bicycle	Rel access car	Relative travel time bicycle	Relative travel time car	Travel time by bus	Address density	Population density	Job density
1	0.742	Coefficient	-0.23396	0.00154			0.32953		0.00126	-0.00183		0.00128					0.00062			
		Std. Error Significance	0.026	0.395			0.061		0.000	0.535		0.056					0.518			
2	0.75	Coefficient	-0.26717	0.00169			0.38448		0.00138			0.00110				-0.02219				
		Std. Error Significance	0.009	0.324			0.017		0.000			0.097				0.433				
3	0.735	Coefficient	-0.19236		-0.00090		0.26229		0.00127			0.00122				-0.01982				
		Std. Error Significance	0.021		0.660		0.041		0.000			0.073				0.494				
4	0.741	Coefficient	-0.26912	0.00164			0.36022		0.00134			0.00118			-0.00332					
		Std. Error Significance	0.010	0.351			0.033		0.000			0.094			0.698					
5	0.755	Coefficient	-0.26375	0.00151			0.32446		0.00132			0.00130								
		Std. Error Significance	0.008	0.365			0.016		0.000			0.034								
6	0.657	Coefficient	-0.06786	-0.00099		0.12716			0.00095			0.00115								
		Std. Error Significance	0.157	0.518		0.329			0.000			0.101								
7	0.757	Coefficient	-0.20924				0.24270		0.00124			0.00137								
		Std. Error Significance	0.005				0.008		0.000			0.025								
8	0.744	Coefficient	-0.21076				0.26584	-0.00002	0.00127			0.00130								
		Std. Error Significance	0.006				0.018	0.692	0.000			0.045								
9	0.784	Coefficient	-0.20333				0.13759		0.00118			0.00158		0.00637						
		Std. Error Significance	0.004				0.176		0.000			0.010		0.104						
10	0.742	Coefficient	-0.21411				0.25517		0.00125			0.00133	0.00415							
		Std. Error Significance	0.008				0.030		0.000			0.046	0.854							
11	0.741	Coefficient	-0.20981				0.24352		0.00124			0.00137								0.00000
		Std. Error Significance	0.007				0.012		0.000			0.031								0.971
12	0.744	Coefficient	-0.20879				0.23642		0.00123		0.00566	0.00139								
		Std. Error Significance	0.006				0.013		0.000		0.707	0.028								
13	0.798	Coefficient	-0.10171						0.00101		0.01792	0.00183						-0.00002		
		Std. Error Significance	0.010						0.000		0.185	0.004						0.002		
14	0.685	Coefficient	-0.06739						0.00111		-0.01077	0.00139							-0.00001	
		Std. Error Significance	0.119						0.000		0.609	0.053							0.076	
15	0.804	Coefficient	-0.11670	0.00161					0.00099		0.02293	0.00185						-0.00002		
		Std. Error Significance	0.006	0.246					0.000		0.108	0.004						0.003		
16	0.779	Coefficient	-0.09473	0.00090					0.00103			0.00166						-0.00002		
		Std. Error Significance	0.019	0.510					0.000			0.009						0.007		
17	0.799	Coefficient	-0.06891						0.00101		0.01774	0.00166				-0.02516		-0.00002		
		Std. Error Significance	0.165						0.000		0.190	0.011				0.323		0.003		
18	0.787	Coefficient	-0.05511						0.00104			0.00151				-0.02564		-0.00002		
		Std. Error Significance	0.262						0.000			0.019				0.327		0.004		

Table H.1 Model parameters of all models estimated without the university and harbour zone