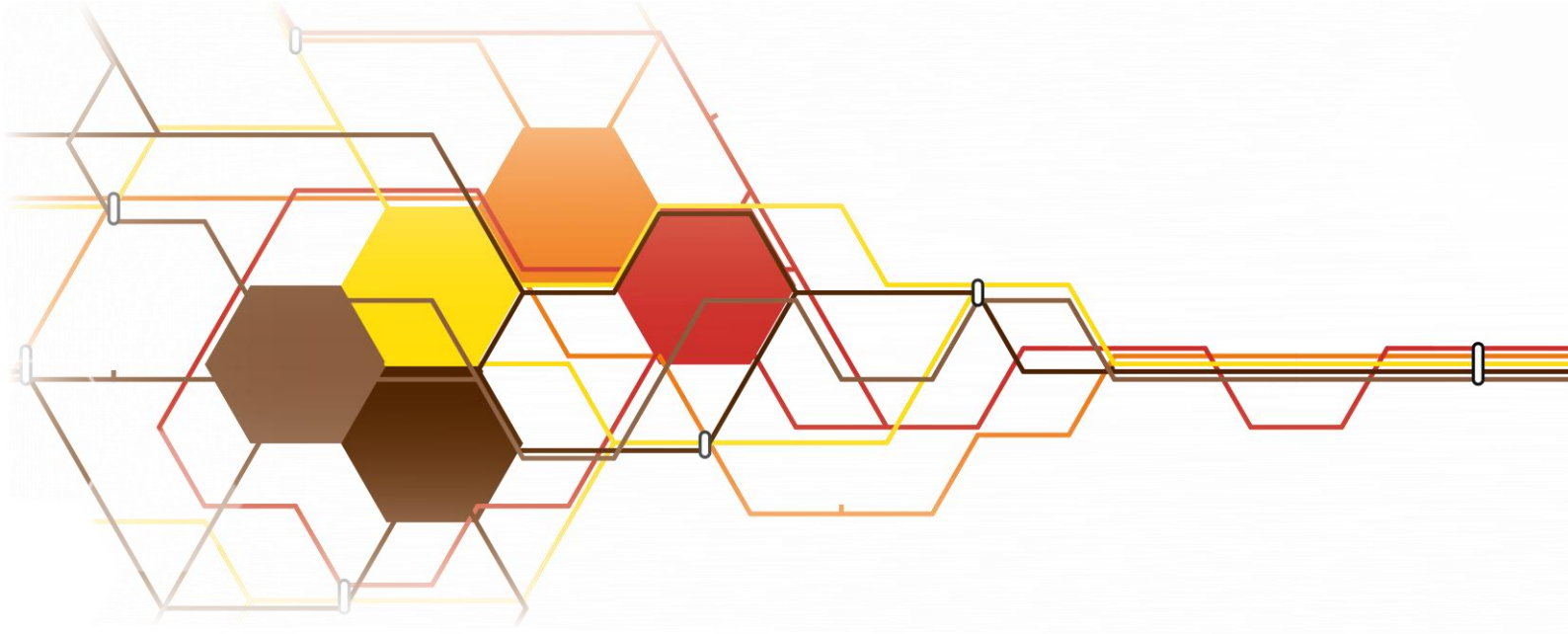
Panteia
Research to Progress

Research voor Beleid | EIM | NEA | I00 | Stratus | IPM

TU Delft Delft University of Technology



The construction of purpose specific OD matrices using public transport smart card data

W. Kuhlman

Colophon

This document contains the master thesis report of Wouter Kuhlman, in fulfilment of the requirements for the degree of Master of Science in Civil Engineering, at Delft University of Technology. The research was performed at Panteia, located in Zoetermeer.

The thesis report has been submitted on Monday October 19th 2015, and will be publically defended on Monday October 26th 2015, at the faculty of Civil Engineering and Geosciences of Delft University of Technology.

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

For any questions regarding the content, please contact:

- Wouter Kuhlman (author) wouterkuhlman@gmail.com or
- Jan Kiel (Panteia) j.kiel@panteia.nl or
- Bert Schepers (Panteia) b.schepers@panteia.nl

Assessment committee

- Prof.dr.ir. E. de Romph – chair (Delft University of Technology)
- Dr.ir. R. van Nes – daily supervisor (Delft University of Technology)
- Drs. J. Kiel – company supervisor (Panteia)
- Dr.ir. M. Kroesen – external supervisor (Delft University of Technology)
- Ir. P.B.L. Wiggenraad – thesis coordinator (Delft University of Technology)

- Drs. B. Schepers – external advisor (Panteia)
- Drs. S. Kieft – external advisor (SRA)
- Drs. N. in 't Veld – external advisor (GVB)

Key words

Smart card data; travel survey data; data enrichment; public transport; OD matrix; travel purpose inference; travel demand modelling

Document statistics

Pages: 149 (127 main text)
Words: 52.049
Chapters: 9
Tables: 27
Figures: 40
Appendices: 4 (external document)

Preface

During my master I got fascinated by the possibilities of smart card data in travel demand studies. This led me to do an internship investigating the potential of a combination of different data sources. Unfortunately, we did not obtain any smart card data during my internship. So, when Panteia and GVB, the public transport operator in Amsterdam, provided the possibility to do my thesis using actual smart card data, I was eager to take it.

One year later, this has resulted in the report that now lies before you. It has been an exciting and challenging journey through the many phases of research. I have learned that, after diving into a subject, I need to come up for air every once in a while. Fortunately, many people supported me during the process.

I am grateful for the opportunity to have performed this research at Panteia. I want to thank my colleagues at Panteia: Bert, Gerben, Ferry, Dick, Yuko, thanks for sharing your experience with me. Thanks Jan, for your guidance and the trips we made to several conferences, at which I was allowed to present my research. I also want to thank the other interns at Panteia for the required distractions at the office.

The creation of this report would not have been possible without the expertise of my graduation committee: Erik de Romph, Rob van Nes, Maarten Kroesen, your feedback and expertise greatly improved the quality of my work. Special thanks to Rob for the sessions in which we structured the many possibilities within this research topic.

I would like to thank the GVB and EBS for making their data available for this research. Special thanks go out to Natalie in 't Veld for her help and enthusiasm during my time at the GVB office in Amsterdam. Also the support from Suzanne Kieft from the City region of Amsterdam is greatly appreciated. I am proud that all these parties committed to this research.

I want to thank my parents for their faith and encouragement and, most of all, my girlfriend Petra, thank you for your love and composure.

Wouter Kuhlman
Utrecht, October 2015

Summary

With the introduction of the smart card as fare collection system in public transport, a new data source emerged. The smart card is able to collect a previously unattainable large sample of travel data by recording all check-in and check-out transactions. When a smart card is the only valid ticketing system, smart card data approximately cover the complete public transport demand. Therefore, smart card data are considered as a rich data source for an abundant amount of topics, predominantly for the description of travel demand.

Problem description

Currently, travel demand models are not able to accurately represent the public transport travel demand. Travel surveys and counts are required to increase the accuracy by means of enrichment and calibration processes. However, collecting this empirical data is expensive. The collection of smart card data is already incorporated in the system and allows for the collection of large data samples at low costs.

Smart card data are very accurate in space and time, as transactions are stored together with their corresponding stop and the time of the transaction. In case the smart card is the only valid ticketing system, like the Dutch OV-chipkaart, the described volumes also accurately represent the volume of the public transport travel demand. Moreover, the continuous data collection allows for longitudinal description of the travel demand.

Yet, the smart card data are also limited regarding their usability in travel demand studies. Smart card data lack information about exact origin and destination locations, where activities are performed. Transactions are recorded at the used stops, hence smart card data do not consider access and egress trip legs to and from the stops. Stop locations are not considered to be stable indicators of the travel demand, as they are not to the actual locations where activities are performed. The used stops relate to the route choice and thus depend on the public transport supply.

Moreover, the interpretability of the travel data is limited due to the passive data collection. Smart card data do not contain information about travellers' motivation to travel, nor personal characteristics and preferences. Because of this lacking information, the interpretation of the travel demand is limited and, consequently, the forecast capabilities of travel demand models using these data are too.

Research motivation

The main objective of this study is the improvement of public transport travel demand forecasts. In order to obtain this objective, we have aimed at increasing the interpretability of smart card data so they can be used as input data for travel demand models.

By enriching smart card data with information about travel purposes and activity locations, smart card data can be used to describe the current travel demand by means of OD matrices. With the high volumes and veracity of smart card data, this results in a more accurate description of the current travel demand. Consequently, travel demand models can provide more accurate forecasts, as the description of current travel demand constitutes the foundation of travel demand forecasts. In

addition, increased interpretability is required for other applications of smart card data in travel demand modelling.

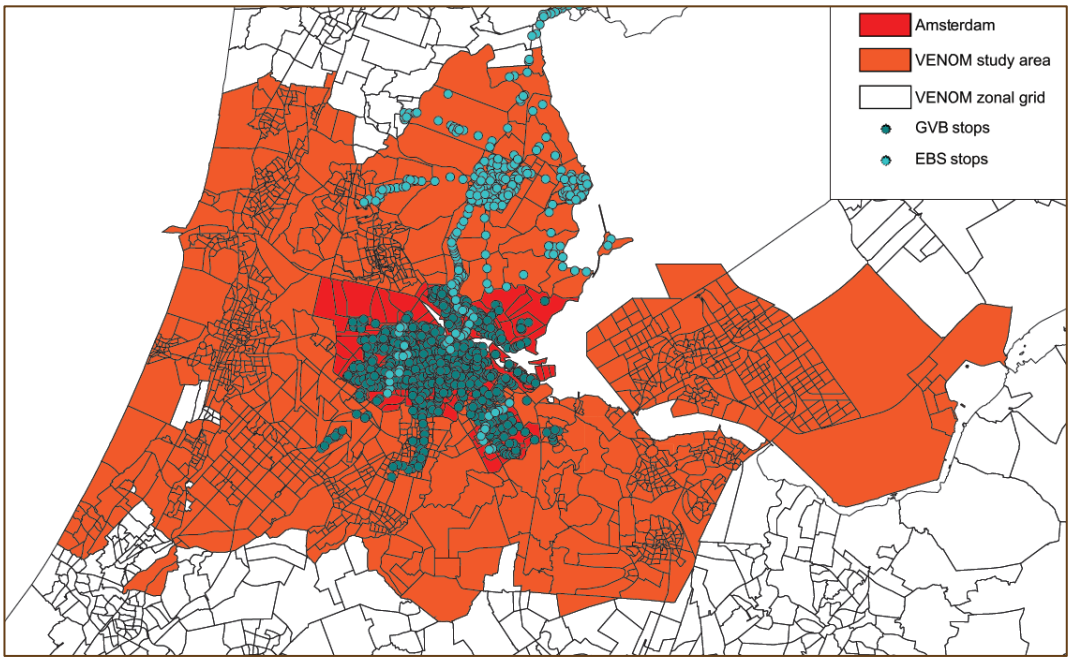
Methodology and data

In order to enrich smart card data, we have developed enrichment models that estimate the lacking information. In order to comply with the highly disaggregated structure of smart card data, travel is represented by individual trips. Three enrichment procedures are applied to every trip, which (1) allocate trips to origin zones, (2) allocate trips to destination zones and (3) infer the travel purpose.

The enrichment models are estimated on data collected by travel surveys. The required information in the survey data consists of the information to be estimated, the dependent variables, and attributes that explain the information to be estimated, the independent variables. Since the models have to be applicable to smart card data, the independent variables are limited to key variables, which are available in both smart card data and survey data.

Data

The smart card data available for this study consisted of the OV-chipkaart data of the year 2014 for two public transport concessions: Amsterdam and Waterland. These concessions are dissimilar, as Amsterdam is highly urbanized and includes bus, tram and metro services, while Waterland is a more rural area to the north of Amsterdam, with only bus services. Data of the national railways and adjacent concessions were not available.



The survey data used for estimation of the enrichment models consist of the WROOV studies, performed yearly between 2003 and 2009. The data for these seven consecutive years are stacked, which results in a dataset of 1.7 million trips with bus and light rail in the Amsterdam and Waterland concessions. The WROOV data include trips by bus and light rail, made with any public transport ticket before implementation of the OV-chipkaart, except student cards and local tickets.

The land-use data available for this study originate from the strategic transport model of the Amsterdam City Region: VENOM. The VENOM model uses a relatively high resolution zonal structure, with corresponding land-use data from the base year 2010.

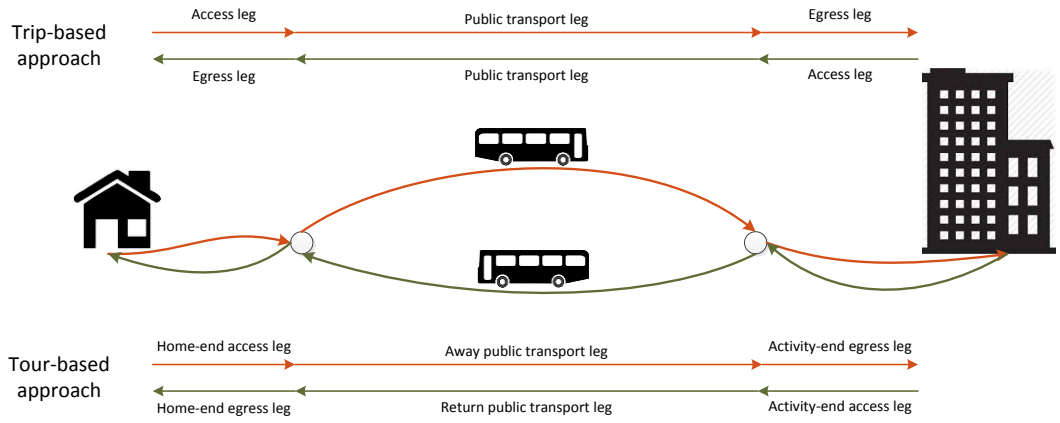
Inevitably, the available data have several limitations. Transfers to and from the train network are not observed, since data from the national railways is not available. The WROOV surveys did not include the student public transport card and local tickets. As a result, the survey data is not suitable to estimate enrichment models for students and tourists. For this reason, we have filtered students (20%) and short term contracts (9%) from the OV-chipkaart data.

Methodology

Literature suggests rule-based methods for the allocation of origins and destinations and the inference of the travel purpose. However, deterministic rules imply crude simplifications. While this may be appropriate for the interpretation of system averages, it is not for the construction of purpose-specific OD matrices. The simplifications result in inaccurate allocation of trip to OD pairs and to wrongly inferred purposes. Therefore, we propose probabilistic methods for the allocation of trips to origin zones, destination zones and the inference of the travel purpose. The enrichment procedures are based on logit allocation models, a technique derived from discrete choice modelling.

We have applied two approaches for the probabilistic enrichment of smart card data: a trip-based approach and a tour-based approach. Trips describe travel between successive activity locations: the origin and destination. The available attributes consist of trip characteristics and land-use characteristics at both trip-ends.

Tours describe travel starting and ending at the home location, via one or several activity locations. The tour-based approach defines trip-ends by their geographical location: the home-end and activity-end, instead of the chronological order in the trip-based approach. This distinction allowed for the application of land-use data related to homes or activities. Regarding the purpose inference, the tour-based approach also included the additional attribute activity duration.



We have compared the probabilistic approaches with a rule-based method at different levels of zonal resolution in order to assess the accuracy increase at each level. In order to do so, we applied the enrichment models to the WROOV data to construct purpose-specific OD matrices for each method. Subsequently, we compared the

resulting matrices with the observed WROOV OD matrix by means of linear regression on the cell values.

In addition, we applied the enrichment models to the OV-chipkaart data and compared the resulting matrices with the WROOV matrices to assess the difference of the described travel patterns between the sources.

Results

Analysis of travel patterns in the WROOV data

Regarding the zonal allocation models, the distribution of access and egress distances proved to differ substantially between the used modes. Especially for metro, travellers cover larger distances to stops on average. Moreover, a distinction is visible between the home-end, where access and egress distances mainly depend on the level of service, and the activity-end, where the access and egress distances depend more on the travel purpose.

Regarding the purpose inference, the distribution of travel purposes in the survey data show large deviations over many key variables, including the departure time, activity duration, travel frequency and ticket types. Age and gender are also correlated to the travel purpose, but not available in OV-chipkaart data and thus not applicable in the enrichment models.

A quantitative comparison of the travel described by the WROOV data and the OV-chipkaart data regarding key variables shows several dissimilarities. Most imbalances are related to the same problem: the overrepresentation of contracts in the WROOV data, resulting in an overestimation of work trips. These dissimilarities could also be caused the time gap between the sources and different data collection methods. Moreover, the WROOV data mainly consist of tours with two trips due to the survey format, where the OV-chipkaart data also contains tours with more than two trips.

Enrichment models

In the trip-based zonal allocation models, similar results were found for the origin inference and the destination inference, due to the high share of tours in the dataset. Three attributes proved to have a significant explanatory value to the origin and destination zones:

- The share of the catchment area in the zone [%];
- The stop density [stops/ha];
- The level of urbanization [addresses/ha].

The share of the catchment area in the zone relates to the nearness of a zone to the used stop in terms of area. It is positively related to the probability of alternatives and has the largest effect on the model fit. The stop density and the level of urbanization relate to the production and attraction of zones. Where the level of urbanization directly relates to potential activity locations, the stop density is an indirect indicator. Although direct indicators are preferable, the stop density was a better and more stable indicator than other direct attributes related to activity locations, such as the number of jobs and the number of student places in a zone.

In the tour-based approach, the same attributes are included in the home zone allocation model. However, the effect of the level of urbanization is substantially larger. On the other end, in the activity zone allocation, the effect of the level of

urbanization is not a stable indicator and thus excluded from the model. As a result, the home zone allocation model has a better fit. The attributes related to the activities, like jobs and student places, did not have a stable influence on the probability of zones.

The trip-based purpose inference models include five attributes, four categorical and one numeric:

- The concessions travelled in;
- The contract duration;
- The travel frequency [tours/week];
- The used modes;
- The departure time.

A comparison of the estimated probabilities and the observed travel purposes in WROOV show that the model is very capable of distinguishing work trips and other trips. Conversely, the identification of education trips and shopping trips is less accurate. This is caused by the similarities between these less occurring purposes and their more frequently observed counterparts, respectively work and other.

The tour-based purpose inference model has a substantially better fit, mainly due to the inclusion of the activity duration. Nonetheless, the distinction between work and education trips is still limited. The inclusion of the activity duration as categorical variable might improve this distinction.

Matrix evaluation

Assessing the matrices constructed with the different approaches, we conclude that the trip-based and tour-based zonal allocation models result in a significantly better representation of the travel demand than a rule-based approach. Especially at a high resolution, like the VENOM zonal grid, the improvement is substantial. The effect reduces at the level of PC3 zones, which indicates that the effect of access and egress legs is relevant for travel demand models with a higher resolution than PC3 zones. This includes nearly all Dutch travel demand models.

Individual matrix cell values, however, still show substantial deviations. The largest deviations are all related to transfers with the train network. Since the train network is highly intertwined with the urban and regional public transport in the Amsterdam region, the unavailability of train data constitutes a limitation to this study.

The comparisons of purpose-specific matrices verify the high accuracy of the inference of the purposes *work* and *other* and the lower accuracy of the inference of *education* and *shopping* purposes. Between the lower scoring purposes, the shopping matrix indicates a better fit than the education matrix. This leads to the conclusion that education trips have more specific spatial patterns, which deviate from commuting patterns, while shopping trips have a similar pattern as trips made for *other* purposes. Consequently, an accurate distinction is possible between compulsory purposes, work and education, and discretionary purposes *shopping* and *other*.

Assessing the differences between the matrices constructed with the different sources, WROOV and OV-chipkaart, we conclude that the described travel demand differs substantially. This can be attributed to differences between the samples, an inevitable

issue when coupling two different data sources, and shifted demand between the data collection periods.

Conclusions

The applied methodology of constructing purpose-specific OD matrices based on smart card data shows great potential. During this research, an operational method has been developed, which results in a more accurate description of public transport travel demand than previously available.

Based on the accuracy of the enrichment models and the constructed OD matrices, we conclude that the tour-based approach is the most accurate enrichment method. Furthermore, the tour-based approach contains a higher level of behavioural richness and thus it is preferable over the rule-based approach and the trip-based approach.

However, restrictions of the available data have resulted in limited applicability and durability of the method. The survey data do not cover students in higher education and international travellers. With additional fine-tuning of and increased availability of smart card data, including all public transport operators in the study area, the method presented in this report can be enhanced to a fully applicable approach and lead to valuable improvements of the quality of public transport demand forecasts.

Recommendations

This study has shown that the probabilistic method of enrichment of smart card data improves the description of public transport travel demand and increases the interpretability of the data. However, due to limitations in the available data the constructed matrices do not represent the complete travel demand. Especially the interaction with train travellers is essential to the description of travel demand by OD matrices. Furthermore, the lacking students and international travellers in the survey data resulted in a substantial gap in the total travel demand.

Therefore, we recommend starting a new, online travel survey in connection with OV-chipkaart data. In order to increase the interpretability of smart card data, periodic travel surveys are still required, as the durability of the method requires periodical updating of the model parameters. With a new survey, a representative sample of travel in the complete public transport system can be selected for estimation of the enrichment models. It is recommended to include the smart card data from all public transport operators in the data in order to identify concession traversing transfers.

Regarding the enrichment methodology, we recommend further research on combining the activity zone allocation and the purpose inference into a single allocation model. The data analysis has shown a correlation between the travel purpose and the access and egress distances at the activity end. Moreover, by combining these models, the effects of purpose-specific land-use characteristics are specifically estimated for their corresponding purpose.

Table of contents

Tables

Figures

Glossary of terms

Access leg	The trip leg between the origin and the boarding stop.
Activity	An act performed outside home, which requires travelling.
Activity-end	The trip-end at the activity side of the trip. In case of an away trip, equal to the destination, in case of a return trip, equal to the origin.
Alighting	Disembarking a public transport vehicle at a stop or station.
Anonymised data	Data that cannot be traced back to an individual.
Automated fare collection (AFC)	A digitalized ticketing system that releases the driver of a public transport vehicle from the task of collecting the fares of travellers.
Boarding	Entering a public transport vehicle at a stop or station
Catchment area	The area around a public transport stop or station from which travellers are drawn.
Check-in	The smart card transaction made when boarding a vehicle (in an open system) or entering a station (in a closed system)
Check-out	The smart card transaction made when alighting a vehicle (in an open system) or leaving a station (in a closed system)
Closed system	A closed public transport smart card system involves smart card readers at the entrances and exits of stations, generally integrated in gates (generally applicable to metro systems).
Concession	A set of public transport services granted to an operator by the regional transport authority.
Destination	The final location of a trip, where the next activity is performed.
Egress leg	The trip leg between the alighting stop and the destination.
Euclidean distance	The distance between two locations measured by a straight line between them.
Evening peak	The peak period of travel demand during the evening. In this research the evening peak is defined from 4 pm to 6 pm.
Fare	The pricing system used for travel by public transport.

Home-end	The trip-end at the home side of the trip. In case of an away trip, equal to the origin. In case of a return trip, equal to the destination.
Key variable	A variable available in both smart card data and survey data, which can be used to compare the sources and transfer information between them.
Land-use characteristics	Characteristics of the functions of land.
Mode	The type of transportations, relating to the infrastructure and vehicles used.
Morning peak	The peak period of travel demand during the morning. In this research the morning peak is defined from 7 am to 9 am.
Non-home-based trip	A trip that does not start at home. (In this study a trip that is not part of a tour is assumed to be non-home-based.)
OD matrix	Describes the number of trips from each Origin zone to each Destination zone, used to describe the travel demand.
Open system	An open public transport smart card system involves smart card readers in the vehicles and does not require gates (generally applicable to bus and tram).
Origin	The start location of a trip.
Smart card	A plastic card with a chip that stores data. In case of public transport smart cards, the chip contains information on the loaded contracts or stored value on the card.
Smart card reader	A device that registers transactions made with the smart card. The reader can read and write data from and to the smart card. In case of public transport smart cards, the reader registers check-in and check-out transactions, as well as loading stored value or contracts.
Time of day (TOD)	A specific period during the day.
Tour	The travel of an individual from the home location, conceivably via activity locations, back to the home location. (In this study, we assume all tours to be home-based.)
Transfer	Change of mode or vehicle during a trip, linking two consecutive trip legs. A transfer can be made between different services at a single stop or consist of a walking leg between different stops.
Travel	The movement of people between different locations.

Travel purpose	The reason for travelling, related to the activity that is performed.
Trip	The travel of an individual from one activity location to another, which may consist of one or more trip legs
Trip leg	Part of a trip covered with a single mode or vehicle, without discontinuations or transfers.
Trip-ends	The two locations between which a trip is made. From a trip perspective the trip-ends are defined chronologically, as the origin and the destination. From a tour perspective trip-ends are defined geographically, as the home-end and the activity-end.
Working days	Monday to Friday
Zone (traffic analysis zone)	A specific area defined by the transport planner, often derived from postal codes. Zones can be defined in different levels of spatial resolution, depending on the

Abbreviations used in this report

AFC	Automated Fare Collection
BTM	Bus, Tram and Metro
LMS	Dutch National Transport Model (Landelijke Model Systeem)
MON/OViN	Dutch Mobility Study (Mobiliteitsonderzoek Nederland/ Onderzoek Verplaatsingen in Nederland)
NRM	Dutch Regional Transport Model (Nederlands Regionaal Model)
NVB	National ticketing system (Nationale Vervoerbewijzen)
OD	Origin-Destination
PC3	3-digit postal code
PC4	4-digit postal code
SRA	City region of Amsterdam (Stadsregio Amsterdam)
TOD	Time-Of-Day
VENOM	Amsterdam Region Transport Model (Verkeerskundig Noordvleugel Model)
WROOV	The Netherlands Public Transport Allocation System (Werkgroep Reizigers Omvang en Omvang Verkopen)

1 Introduction

Since 2009, a new ticketing system was introduced in the Dutch public transport sector: the OV-chipkaart. This smart card system is currently employed as an Automatic Fare Collection (AFC) system by all public transport operators in the Netherlands. The fare collection works as follows. When boarding, travellers present their smart card to a reader, which generates a check-in transaction. Subsequently, the smart card is again presented to the reader at alighting, generating the check-out transaction. Fares are calculated based on the times and locations of these two transactions. Over the year 2014, the smart card system operator Translink registered 2.2 billion[1] transactions (Translink, 2015), which were all stored in the central back office.

The OV-chipkaart replaced the National Ticketing System (NVB), together with the complementing survey studies of The Netherlands Public Transport Allocation System (WROOV). The NVB contained the main ticket types in public transport, which were available throughout the Netherlands, and allowed travellers to travel with all public transport operators. The WROOV studies consisted of surveys that were used to allocate the revenues of the NVB to the operators and public transport authorities. The survey involved questions about the use of public transport based on ticket sales. Besides the information required for the fare box allocation, the survey also generated data on personal characteristics, the origins and destinations of travellers and the travel purpose. Consequently, the WROOV surveys were the primary data source of travel patterns regarding bus and light rail until 2009, the year the OV-chipkaart was introduced countywide.

This transition from conventional data collection towards new digital sources provides both opportunities, as well as challenges, for the use of public transport travel data. Digital data collection offers the possibility to handle large amounts of data at lower costs, resulting in high veracity on the number of travellers. However, since the data collection is passive, no additional information is collected on the characteristics and preferences of travellers. Combining the high veracity of smart card data with information derived from survey data might provide new opportunities for the application of public transport travel data.

1.1 Problem description

Public transport smart card data contain valuable information for operators, planners and authorities, and therefore they are applied in numerous fields. Some examples are transport performance monitoring and analysis, travel demand modelling, route choice modelling, advertising and even detection of contagious outbreaks. This study focusses on the utilization of smart card data for OD matrix construction, as input for travel demand models. Travel demand forecasts are essential information for authorities that take strategic decisions on investments in transport supply. Travel demand models are the primary tool to render these forecasts, therefore accuracy and credibility of these models are key in the decision making process (Ortúzar & Willumsen, 2011). These performance statistics depend on the quality of the input data. However, collecting this data by means of surveys and counts is expensive and

[1] Besides check-in and check-out transactions, the 2.2 billion transactions also include transactions recharging the stored value on smart cards.

sample sizes, and thus the quality of the results, is under pressure in these times of budget cuts.

The formulation of Origin-Destination (OD) matrices is a primary method for describing the travel demand and therefore it plays a prominent role in several modelling techniques (Ortúzar & Willumsen, 2011). These matrices contain the number of trips from every Origin zone to every Destination zone. In practice, the synthetic OD matrices generated by *trip generation* and *trip distribution* models often do not represent the current travel demand as it is observed in reality. This deviation can be associated with two issues:

1) The model does not take into account the daily variation of travel, but generates the travel of an average working day;
2) The model does not capture all factors that influence travel behaviour.

The first issue affects both aggregate and disaggregate modelling approaches and traditional data collection techniques do not provide the tools to compensate for it due to the limited sample sizes (Ortúzar & Willumsen, 2011). Regarding the second issue, correction is sought by means of matrix calibration. This process alters the matrix values in order to improve the fit of the matrix to calibration data, traditionally collected by means of counts or OD surveys. However, the influence of the calibration process should not be too large. The input data may not represent the average working day and large alterations can compromise the model consistency. Ortúzar & Willumsen (2011) state that the objective of the calibration process should not be to replicate the observations, but to estimate a matrix that captures their main features.

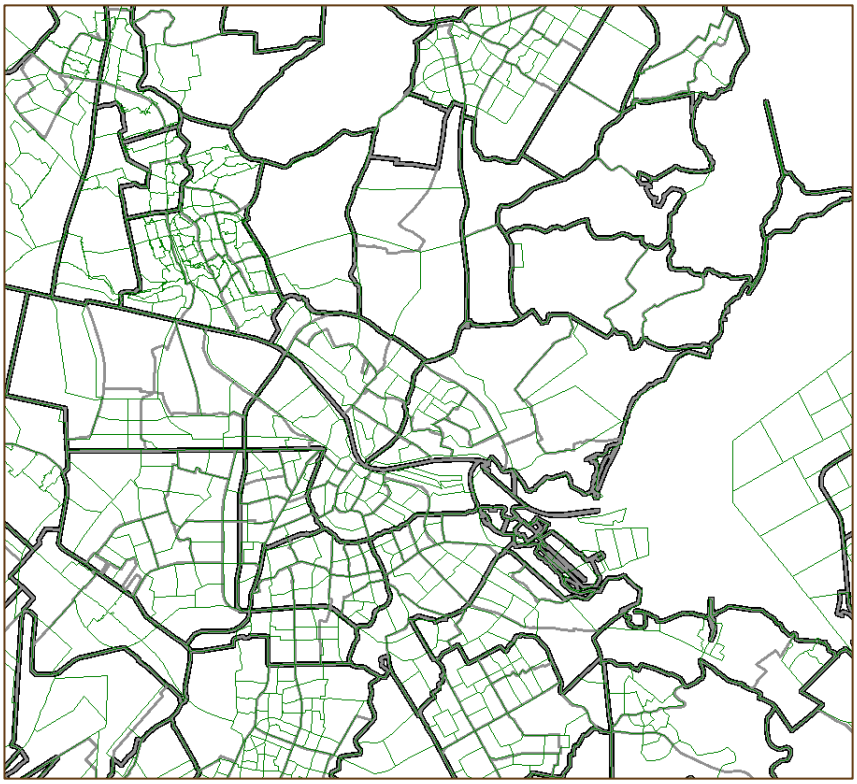


Figure 1: Amsterdam in three different zonal resolutions

Public transport smart cards provide continuous, disaggregate data that is very detailed in both time and space. Boarding and alighting times and locations are

recorded in detail. As they are often used to determine the price of the travel, they are the primary goal of collecting the data. However, the information that these features provide is still limited with regard to OD matrix construction. In order to specifically implement the effects of policy measures and socio-economic forecasts in the model, OD matrices are split into different categories that relate to the travel behaviour of travellers. These characteristics relate to the choices a traveller makes, and are affected by policy measures and changes in socio-economic data (Rijkswaterstaat, 2012). Therefore, OD matrices are categorized by:

- Travel purpose (why travel?);
- Mode (how to travel?);
- Time of Day (TOD) (At what time to travel?).

Considering public transport smart card data, the travel purpose is not observed. The modes are limited to public transport modes and can be derived from the data. The total number of trips made and the times of travelling are also directly available from the data, as well as the locations of the used stops. Hence, the specification of stop-based matrices by mode and by time of day is uncomplicated, but the distinction by travel purpose

The locations of the used stops are, however, not the equal to the origin and destination. The origin and destination of a trip are the locations where activities have been performed, for example working, but also staying at home. In order to arrive at the boarding stop, a traveller has to make an access trip leg. On the other end of the trip an egress trip leg leads from the alighting stop to the destination. Access and egress trip legs are usually short and mostly covered on foot or by bike (Goeverden). Therefore, the stop location can be a close approximation of the origin or destination. Yet, the use of stop locations can lead to wrongly allocated zones since access and egress trip legs can cross zonal borders, especially with a high resolution zonal grid.

Furthermore, the travel purpose is not available in smart card data. The data collection is passive, and therefore traveller characteristics are unavailable. For privacy reasons, is it prohibited by law to couple travel data to personal data[2]. Consequently, for the construction of OD matrices, trips observed by smart card data lack three essential pieces of information, the information we aim to add:

1. The origin of the trip;
2. The destination of the trips;
3. The travel purpose.

Where smart card data provide large numbers of observations, survey data can provide more insight in the background of the traveller. Surveys can include questions on the travels made, providing information on the origins and destinations, used modes and TOD, but also questions regarding the traveller, which provide information on the gender, age and the travel purpose. Hence, survey data contains all the required information for the construction of OD matrices.

However, to construct OD matrices based on surveys, the sample size should be sufficiently large to capture the main features of the travel patterns in the system. For high resolution zonal grids, this requires large sample sizes, which increase the expenses substantially.

[2] [2] Wet Bescherming Persoonsgegevens (WBP, BWBR0011468)

Where smart card data can provide high veracity on the number of trips, used modes and TOD, survey data can provide information on the access and egress patterns and travel purposes. Together these sources contain the information to construct OD matrices with high veracity, at low costs. Therefore, this study focusses on coupling of these two data sources.

1.2 Research motivation

We have distinguished one main objective of this study and two supporting objectives. This paragraph describes the motivation for these objectives, with the matching aims and specific goals to achieve them.

1.2.1 Objectives, aims and goals

The main objective of this study is to improve the public transport travel demand forecasts, and thereby enhance the decision making process of authorities concerning investments in public transport. We aimed to achieve this by starting at the front side of transport models and improve the quality of the input data that describe the current travel demand by means of OD matrices.

Other options to enhance travel demand forecasts with smart card data are to calibrate transport models with the large amount of observations or to create new modelling techniques based on this new source. We have focussed on the description of the current travel demand as this can be applied in current methodologies and it provides the essential insight of the competences of smart card data. Thereby, this study provides a foundation for research in the application for model calibration and development of new modelling techniques, which can be encouraging follow-up studies.

The implementation of new data sources is not only a way to improve travel demand forecasts, but it is also an objective in itself. Globally, more and more developments are information driven. This trend has also penetrated into the transport market. Travellers want to know their route options and travel times before departure, so they can choose their route and departure time deliberately, based on accurate information. Transport modelling needs to keep up with these developments, and the information producing travellers provide several options to do so.

While modelling techniques are usually very much settled over time in order to preserve consistency, new data sources need to be adopted to comply with current developments. In addition to improving the modelling accuracy, this is also important for the replacement of traditional data sources with declining efficacy. New data sources are, however, not direct replacements of traditional data sources. This study aimed to increase insight in the differences between smart card data and survey data and the requirements for application in OD matrix construction.

The implementation of new data sources is in turn supported by the objective to increase insight in public transport travel patterns. By relating the travel purpose and access and egress trip legs to other travel characteristics we have gained insight in the behaviour related to these lacking pieces of information in smart card data.

1.2.2 Relevant actors

Different types of actors can benefit from this research. Primarily, users of strategic transport models benefit from an increased quality of the OD matrices. These users

consist of authorities responsible for procurement of PT concessions and municipalities. Improved modelling of travel with public transport will help them to make better choices regarding investments in transport supply and also increase the acceptance of model results by involved parties. In addition, PT operators will benefit from this study with increased perception of the travel behaviour in their system as well as an improved long-term forecast of the travel demand.

Furthermore, transport consultancies and research institutes can benefit from the newly developed modelling techniques and acquired knowledge on the combination of travel patterns in public transport. The conclusions and recommendations on combining sources might be relevant for future research, not only on the use of smart card data, but also on the implementation of other Big Data sources in transport modelling.

1.3 Research questions and approach

Directing the research motivation to the problem at hand, one main research question has been formulated. In order to structure the research process, the main question is divided into four sub questions, targeting the distinct aspects of this study.

1.3.1 Research questions

To achieve the main objective, this study has answered the following research questions:

"To what extent can the travel purpose, origins and destinations of public transport trips derived from smart card data be inferred based on information from survey data, in order to construct purpose-specific OD matrices suitable as base matrices in transport models?"

Survey data contain all the required information for the construction of OD matrices per purpose, mode and TOD. Hence, we are able to estimate the relations between the information that is lacking in smart card data and characteristics that are available, based on survey data. These relations are represented by enrichment models, which can project the relations found in survey data onto smart card data. Consequently, we combine the high veracity of smart card data on the number of trips, with the lacking information from survey data.

Only attributes that are available in both sources, the key variables, can be used to transfer information from survey data to smart card data. Hence, the enrichment models can only apply key variables as independent variables to estimate the lacking information.

First, we aim to identify attributes that are correlated to the information we want to add to the OV-chipkaart data by means of an analysis onto WROOV data. Second, we aim to assess the appropriateness of these attributes as medium of transferring information by comparing the data sources both qualitatively and quantitatively. Based on this assessment, the datasets are adapted in order to match their coverage of the system. In addition, attributes are selected for implementation in the enrichment models. Third, we aim to optimize enrichment models that can be applied for the construction of purpose-specific OD matrices, based on the appropriate attributes with potential explanatory value of the lacking information. Fourth, and finally, we aim to evaluate the constructed OD matrices by comparing matrices

constructed by different modelling approaches in order to assess their quality and value.

In order to represent these different subjects within the research, the main research question is divided into the following sub-questions:

1. Which travel characteristics are correlated to the information to be added to OV-chipkaart data and to what extent?
   1.1 Which travel characteristics, and to what extent, are correlated to the access and egress trip legs to and from public transport stops?
   1.2 Which travel characteristics, and to what extent, are correlated to the travel purpose?
2. How do the data sources OV-chipkaart and WROOV compare to each other?
   2.1 How do the data sources OV-chipkaart and WROOV compare to each other qualitatively, concerning data collection method, available information, target populations and coverage of the transport system?
   2.2 Which attributes are key variables and how can these be represented in order to construct comparative data sets from OV-chipkaart and WROOV?
   2.3 How do the data sources compare quantitatively, regarding the travel they describe?
3. To what extent can information, lacking in the OV-chipkaart data, be inferred based on WROOV data, in order to construct purpose specific OD matrices?
   3.1 To what extent can origins and destinations of trips derived from smart card data be inferred in order to convert boarding-alighting matrices, based on stops, to OD matrices, based on zones?
   3.2 To what extent can travel purposes of trips derived from smart card data be inferred in order to distinguish base matrices by purpose?
4. How do base matrices created by different methods compare to each other?
   4.1 To what extent do the trip-based and tour-based approaches result in different matrices for the total average working day, matrices per time of day and matrices per purpose?
   4.2 At what level of resolution do these differences appear?
   4.3 How can continuity of this method of OD matrix construction be attained regarding the required data sources?

1.3.2 Approach

The foundation of this study was laid during the literature study on the applications of smart card data for travel demand forecasts. The literature study provided possible methodologies to couple the data sources as well as indications of which attributes could have explanatory value.

In order to describe the travel purpose and the access and egress distances based on other travel characteristics, an analysis of the survey data was performed. By means of this analysis, and with input from the literature study, we answered sub-question 1 and learned which characteristics could possibly have an explanatory value in the in the coupling of information to smart card data.

Subsequently, the data sources were compared in order to derive the transferability of information. In order to couple the data sources, key variables were identified. These variables are available in both data sources and information can be added to smart card data via these attributes. In order to determine the appropriateness of these variables as medium for transferring information, their patterns in both sources were

compared. Combined with a qualitative comparison of the data sources, we answered sub-question 2. This leads to the selection of key variables to start the model estimations.

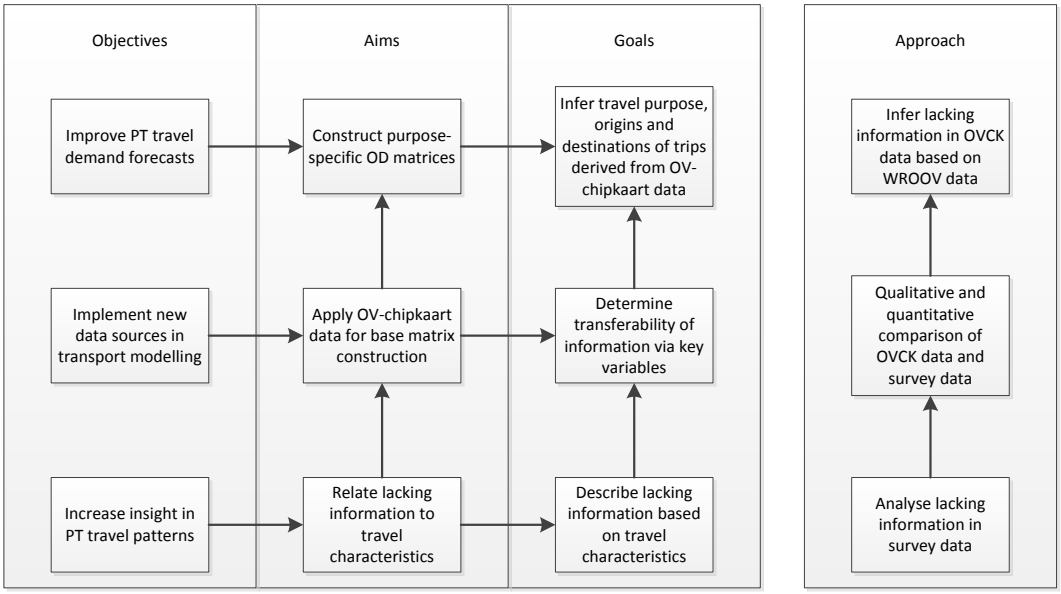


Figure 2: Research objectives, aims and goals

Using these key variables as attributes, we estimated zonal allocation models and purpose inference models. These have been applied to enrich smart card data with the information required to construct purpose specific OD matrices. During the model estimation phase, the explanatory variables were further selected based on the model performance. Three different approaches have been applied, with different levels of enhancement. These models were validated on WROOV data, to investigate their accuracy and answer sub-question 3.

Finally, we have evaluated the differences between the model approaches by comparing their resulting OD matrices. The evaluation consists of comparisons at different levels of spatial resolution, as well as time resolution and purpose specification. These comparisons have provided insight in the appropriate modelling approaches for specific OD matrix qualifications, and thereby answered sub-question 4.

1.4 Definitions used in this report

The perception of several definitions used in this report is essential to grasp the subtle distinctions that have played a significant role in this study. Therefore, we present the most essential terms here. A complete glossary of terms can be found at the front of the report.

Firstly, there are several ways to describe travels, which are closely linked and should not be confound. The principal unit of travel used in this study is a *trip*. A trip is made from one activity location, the origin, to the next activity location, the destination. OD matrices are based on trips, as these are directly related to the performance of the transport system. Zooming in, a trip can consist of several *trip legs*, which relate to a certain mode or vehicle that is used. Consecutive trip legs within a trip are connected by transfers.

Zooming out, research shows that consecutive trips made by a traveller, are often related. Therefore, an alternative approach in describing travel was introduced, based on *tours*. Tours describe the travel from one location up until return at that location. Tours can be classified as home-based, thus starting and ending at home, or non-home-based, starting and ending at, for example, the office. Several activities can be performed within one tour, increasing the number of trips within the tour. Hence, a tour with one trip relates to a tour without activities. A tour with two trips relates to a tour with one activity, et cetera. The distinction between transfers, that connect trip legs, and activities, that connect trips, is that a transfer only serves the purpose of getting onto the next vehicle (Devillaine, Munizaga, & Trépanier, 2012).

The trip-based and tour-based approach result in different taxonomies of the trip ends. The trip-based definitions of *origin* and *destination* are applied in the OD matrix. Conversely, applying the tour-based approach, these are classified as the *home-end* and *activity end*. By using this definition, trip ends are categorized by location instead of departure or arrival. Compared to the trip-based approach, this allows for the anticipation of specific characteristics at certain locations and ensures consistency between trips. This distinction has proven to be rather important in the allocation of trip ends to zones.

1.5 Report structure

The remainder of this thesis report is structured as follows.

- Chapter 2 covers the results from the literature review. Here, relevant references from literature are discussed regarding the use of PT smart card data in travel demand modelling. In addition, this chapter lists methodologies applied in other studies, used to handle similar questions as the ones covered in this study;
- Chapter 3 presents the methodology used for answering the research questions and the data sources employed;
- Chapter 4 contains a description of the case study used for answering the research questions. This chapter covers the availability of data for this study and the complementary boundaries, as well as the relation between these boundaries and the generalizability of this study;
- Chapter 5 comprises the results from the WROOV data analysis, regarding travel characteristics related to access and egress distances and travel characteristics related to the travel purpose. Furthermore, the data sources WROOV and OV-chipkaart are compared, both qualitatively and quantitatively;
- Chapter 6 covers the process and the final results of the estimation of zonal allocation models and purpose inference models. Here, the interpretation, stability and generalizability of the final models are discussed;
- Chapter 7 encompasses the evaluation: the comparison of the OD matrices constructed by different modelling approaches;
- Chapter 8 lists the conclusions from this study and discusses the implications of the findings. The conclusions are divided in the answers of the research questions and additional conclusions;
- Chapter 9, finally, contains the recommendations to relevant actors based on the conclusions. The recommendations are categorizes by the utilization of the results and recommendations regarding further research.

Table 1: Locations of sub-questions in the report

Sub-question	*Paragraph*	*page*
1. Which travel characteristics are correlated to the information to be added to OV-chipkaart data and to what extent?	5.1 / 5.2	51 / 54
2. How do the data sources OV-chipkaart and WROOV compare to each other?	3.4 / 5.4	30 / 60
3. To what extent can information, lacking in the OV-chipkaart data, be inferred based on WROOV data, in order to construct purpose specific OD matrices?	6.2 / 6.3	68 / 82
4. How do base matrices created by different methods compare to each other?	7.2	99

2 Literature review

This chapter contains the results of the literature study that has been performed throughout the research. The literature study functions as the foundation of the research and has four objectives:

- *Providing background information*: ensuring comprehension of the possibilities and complexities of smart card data and their utilization;
- *Listing reference studies*: placing this research in context of other studies, extracting relevant information and methodologies and identifying gaps in the existing literature;
- *Discovering methodology options*: determining feasible options to answer the research questions;
- *Supplying input for the assessment framework*: formulating a comprehensive and consistent framework of assessment criteria for the methodology choice.

The chapter is structured as follows. First, the smart card data are introduced (section 2.1). This paragraph describes the implementation of the smart card in PT, the data structure and the potential of smart card data for travel demand modelling. Second, methodologies to construct OD matrices from smart card data are discussed (section 2.2). This leads to an indication of possible solutions to answer research question 3.1. Third, techniques for travel purpose inference are discussed (section 2.3). This leads to an indication of possible solutions to answer research question 3.2. The chapter culminates with conclusions on the discovered literature and implications for this study (section 2.4), supplying an overview of possible methodologies and assessment criteria.

2.1 Introduction to smart card data

Smart cards have become a popular means of fare collection in PT systems. Renowned examples are the Octopus Card in Hong Kong and the Oyster card in London, but there are many more smart cards systems emerging around the world. Without going into detail on the smart card technology, this paragraph describes the data that is collected through smart card systems used in PT.

2.1.1 Introduction of the smartcard in public transport

Early contributors to the research on the use of public transport smart card data did not directly focus on the possibilities for transport demand modelling. Blythe was primarily interested in the interoperability of smart cards between operators across the UK and additional advantages for the operability of public transport (Blythe & Holland, Integrated ticketing - Smart cards in transport, 1998). These advantages consist of less labour-intensive revenue collection, reduced boarding times and increased security (Blythe, Improving public transport ticketing through smart cards, 2004). Dinant & Keuleers (2004) investigated the use of smart card data from a data protection perspective. They identify the privacy concerns that emerge with cross-profiling of databases and present several cryptographic solutions to prevent the possibility of cross-profiling. Hence, enriching smart card data by cross-profiling is not considered as a desirable solution to this research into travel demand modelling. Moreover, the Dutch law on protection of personal information[3] prohibits the use of personal data for purposes other than the ones specified in advance.

[3] Wet Bescherming Persoonsgegevens (WBP, BWBR0011468)

2.1.2 Structure of smart card travel data

The characteristics of smart card systems in PT can differ in several ways. The universal characteristic is the use for automated fare collection (AFC). However, the recording of relevant data for travel demand studies with these AFC systems depends on several specifications. There are six main differences in the description of smart card data structure observed in literature:

1. The smart card system can cover one or several modes of PT, where differences in the structure of collected data might occur (Seaborn, Attanucci, & Wilson, 2009). A higher share of smart card deployment in the total PT system, results in a more integral representation of the travel behaviour (Cui, Wilson, & Attanucci, 2006).
1. The availability of other ticket types besides the smart card will result in an incomplete, and possibly biased, representation of travel behaviour. Moreover, smart cards are not necessarily equal to unique travellers, since travellers can have more than one card or share their card (Morency, Trepanier, & Agard, 2007) (Robinson, Narayanan, Toh, & Pereira, 2014);
2. Different fare policies result in different transaction requirements. Flat fares only require one transaction (check-in) that reduces the stored value on the card with the fare for one ride. Alternatively, distance based fares require two transactions: when boarding and when alighting (check-in and check-out), to determine the distance travelled. Distance based fares therefore collect more relevant information for travel demand studies, including the alighting location, distance travelled and travel time;
3. The availability of stop or station information where the transaction took place depends on the placement of the smart card reader. Smart card readers can be placed in vehicles or at stations, depending on the transport system. Readers placed at stations are usual for train and metro systems and are generally placed at the station entrances and exits. Since these readers are stationary, they can be automatically coupled to location information. Readers in bus and tram systems are usually placed in the vehicle, thus not stationary. Hence, the availability of stop information depends on the integration with an Automated Vehicle Location (AVL) system, which records the GPS coordinates of the vehicle at the time of the transaction. For distance based fares, the availability of an AVL system is required, but it is also applicable for flat rate fares. In most systems, the AFC and the AVL are one integrated system, assigning every check-in and check-out transaction to a stop or station. If AFC and AVL are separated, stops are not directly recorded in the data, but can still be derived by additional data processing (Nassir, Khani, Lee, Noh, & Hickman, 2011) (Gordon, Koutsopoulos, Wilson, & Attanucci, 2013) (Liao & Liu, 2010) (Zhao, Rahbee, & Wilson, 2007);
4. The identification of transfers can be automated in the data collection, indirectly traceable or completely absent in the data. This depends on the fare policy for trips with several rides and the smart card software. The automated identification of transfers might not always match the definition of a transfer used in this study (see paragraph 2.2.2);
5. The quality of the data is influenced by both system errors and user errors. System errors can include software bugs, hacks, erroneous input data of the network or fare structure, broken hardware and communication malfunctions. User errors can include failure to check-in or check-out and untimely check-outs. Both cases can be either deliberate or not (Robinson, Narayanan, Toh, & Pereira, 2014) (Chu & Chapleau, Imputation techniques for missing fields and implausible values in public transit smart card data, 2007). Despite these possible errors, the overall quality of smart card is high, certainly compared to survey data. In addition, correction for

some of these errors is possible by means of rule-based inference. (Chu & Chapleau, Imputation techniques for missing fields and implausible values in public transit smart card data, 2007)

Smart cards are often personal cards, so technically the smart card data could also contain personal data. Although collected by purchase transactions, personal data are usually not combined with smart card data due to privacy concerns. One study, however, did have access to billing addresses. Utsunomiya et al. (2006) used the data from the Chicago Card to study the access and egress distances to stops.

2.1.3 Potential for travel demand modelling

Bagchi & White (2005) pioneered in the potential of smart card data for travel demand modelling. The authors provide a insightful overview of possibilities and constraints of smart card data and compare the smart card to traditional data sources. They define a conceptual framework for travel behaviour analysis, which they apply to the problem of transport turnover, also referred to as churn. This framework consists of three generic dimensions in which travel behaviour takes place: time, space and structure (see Figure 3). In order to analyse travel behaviour, boundaries for these dimensions need to be determined. Spatial boundaries can be defined by administrative areas, like postal codes, or by service areas of public transport operators. Regarding time boundaries, the authors suggest a period of one year, to balance effects of seasonal variation in behaviour. The structure can be constrained by transport operators, modes or routes.

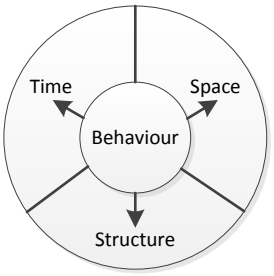


Figure 3: The primary dimensions of travel behaviour (Bagchi & White, 2005)

Analysing the potential of smart card data, the authors encounter essential issues that complicate the use for travel demand modelling. Smart card data do not indicate the exact origins and destinations on the level of street addresses, nor the travel purpose. Furthermore, the authors also acknowledge the problems concerning definitions. Smart cards record passengers boarding and in some systems also alighting, which are not equal to passenger-trips. The UK National Travel Survey defines trips as one way journeys from one activity to another, which may consist of trip legs. Trips can be constructed by data processing, for example by assuming a transfer in case of an alighting followed by a boarding within a certain time interval. This accentuates the importance of concept definitions and rule-based processing, as the interval is often chosen arbitrarily. Another issue is the generalizability of the group of smart card users, which might not be typical for the whole population when several ticket types are in operation.

On the other hand, smart card data also have substantial advantages over survey data. Most surveys are momentary representations of behaviour, while counts, of both

ticket sales and passengers, cannot be traced back to individuals. The personal and continuous characteristics of smart card data are considered to be a valuable enhancement in the examination of travel behaviour, even though smart cards are not necessarily equivalent to individuals. Bagchi and White conclude that smart card data have a large potential due to the large sample size and long periods of coverage, but smart card data alone are not sufficient for travel demand modelling. They suggest a routine survey to verify and complement the smart card data with information on travel purpose, origins and destinations.

Building on the fundamental analysis by Bagchi & White (2005), Pelletier et al. (2011) propose deliberate research topics to employ the potential of smart card data for travel demand modelling. The authors constructed a comprehensive literature review of the use of smart card data for transport planners up to 2011. They differentiated a wide range of studies into operational, tactical and strategic planning. They conclude that smart card data can be useful for both researchers and transport planners, and list challenges to overcome in order to use the full potential of smart card data in transport planning. The authors state that, from a researchers perspective, this new data source requires new modelling approaches that are fit for such a detailed level of resolution. A primary issue is linking socio-demographic data to anonymous smart card data, while complying with privacy regulations. Furthermore, the continuous data will encourage new methods of data analysis for longitudinal studies. The authors end with the notion that the smart cards will be commonly used in public transport, ensuing research on travel behaviour. And indeed, much research has been done since 2011.

Concluding the research from Bagchi & White and Pelletier et al., there are two main issues to solve in order to employ the full potential of smart card data in travel demand modelling:

1. Distinguishing the actual origins and destinations of the trips observed in smart card data;
2. Estimating the travel purpose;
3. Adapting modelling techniques in order to utilize the continuous character of smart card data.

The first two issues are the main topics of this study. Literature on these issues is discussed in the following paragraphs. The third issue is another subject, which could be dealt with in a follow-up study.

2.2 Smart card data for OD matrix construction

In order to take up the first issue of this study, this paragraph discusses the literature on the construction of OD matrices from smart card data. The formulation of OD matrices is a primary method for describing the travel demand (Ortúzar & Willumsen, 2011), and so, this has been a recurring topic in the research on smart card data. As described in the previous paragraph, smart card systems and their collected data can differ substantially.

Depending on the specifications of the smart card system, several procedures have to be performed in order to convert the smart card data into OD matrices. First, if the alighting stop is not recorded, it has to be inferred. Successively, if the used stops are known, the trip legs have to be converted into trips by identification of transfers. Transfers within the described system can be derived through data processing. On the

other hand, transfers to other systems that are not included in the available data cause another problem. In the final step, the stop-based matrices that are derived from smart card data can be converted into OD matrices by inferring the actual origins and destinations of trips.

2.2.1 Alighting stop estimation

In smart card systems with flat rate fares, the alighting stop is not recorded in the data. Consequently, the first step in constructing a stop-based OD matrix is the inference of the alighting stop. Several studies describe the procedure of estimating the alighting stop based on rule-based processing. The fundamental processing rules are (Cui, Wilson, & Attanucci, 2006) (Barry, Freimer, & Slavin, 2009) (Wang, Attanucci, & Wilson, 2011):

1. The alighting stop of a trip leg is the nearest stop to the boarding stop of the next trip leg that day;
2. The last alighting stop of the day equals the first boarding stop of that day.

These rules are based on the assumptions that 1) travellers will return to their alighting stop after their activity, since the walking distances are small, and 2) travellers start and end their travels at home.

In addition to these elementary rules, several refining rules have been applied. The search procedure for the first rule is delimited to stops on the used line, in the right direction. Constraints can be set on the maximum distance and the travel time between alighting stop and boarding stop (Zhao, Rahbee, & Wilson, 2007) (Munizaga & Palma, 2012). If the second rule does not result in a valid alighting stop, the search procedure can be extended to the next day, as tours may cross the day border (Munizaga, Devillaine, Navarrete, & Silva, 2014). Another option to extend the search for an alighting stop is the examination of regularity in trips and estimate the alighting stop based on travel patterns (Trépanier, Tranchant, & Chapleau, 2007) (Jun & Dongyuan, 2013).

Where Barry et al. (2009) find 90% valid alighting stops for the New York metro system, Trépanier et al. (2007) report a success rate of 66%. However, the alighting stops might be valid, but still wrongly estimated since there is no direct validation possible. The need for validation of smartcard data is therefore generally acknowledged and has been studied using survey data. The results of various studies differ considerably: Munizaga et al. (2014) find a correct estimation of the alighting stop in 84% of the cases, where Wang et al. (2011) report only 60% of correctly estimated alighting stops.

2.2.2 Identification of transfers within the described structure

The travel demand we want to describe consists of the number of trips between activities, since activities form the reasons to travel to their corresponding location. Therefore, OD matrices are trip based, where one trip can consist of several trip legs (see paragraph 1.4 for the definitions). Transfers are not activities in this regard, as they only serve the purpose of reaching the destination (Devillaine, Munizaga, & Trépanier, 2012). The smart card does not record trips, but trip legs.[4] In order to

[4] In closed systems, with stationary smart card readers at the entrances and exits of stations, transfers do not require additional check-in and check-out transactions. Therefore, consecutive trip legs within a closed system are considered as one trip leg in the data. Reddy et al. (2009) derive the number of actual trip legs in a closed system based on a survey.

derive trips, transfers need to be identified between consecutive trip legs that are part of the same trip.

While some smart card systems automatically record transfers in the data, transfers are not actually observed. What a traveller does between alighting and a consecutive boarding is unknown. It could be the traveller is walking to another stop or waiting for a connecting ride, but it is also possible that a short activity takes place. The identification involves rule-based processing, usually by means of a time constraint. Therefore, automated identification of transfers is not necessarily the right identification (Nassir, Khani, Lee, Noh, & Hickman, 2011) (Robinson, Narayanan, Toh, & Pereira, 2014). The specified time interval that an operator uses for fare calculations, in case of reduced fares for transfer trips, might even be a cost-based incentive for short activities (Jang, 2010).

For systems without automated recording of transfers, similar rule-based processing methods are applied to derive transfers and combine trip legs into trips. From Bagchi and White (2005), we know constraints can be set on time, space and structure. Constraints on these dimensions are frequently identified as rules to infer transfers within the transport structure described by the smart card data. The proposed constraints, however, differ substantially over the studies found in literature. A range of time constraints is used:

- 30 minutes (Bagchi & White, 2005) (Munizaga & Palma, 2012) (Devillaine, Munizaga, & Trépanier, 2012);
- 35 minutes (Nijënstein & Bussink, 2014);
- 60 minutes (Chakirov & Erath, 2012);
- 90 minutes (Hofmann & O'Mahony, 2005) (Hofmann, Wilson, & White, 2009);
- 120 minutes (Utsunomiya, Attanucci, & Wilson, 2006).

If an operator uses a flat fare structure, the alighting time is not available and the on-board time should be included in the time constraint. This results in different time constraints for transfers between different modes (Seaborn, Attanucci, & Wilson, 2009). Transfer times that lie in between the aforementioned time constraints can be ambiguous. Variance between average transfer times in different networks is observed: in Seoul 80% of the transfers take less than 10 minutes (Jang, 2010), while in Gatineau this share is reached at 18 minutes (Chu & Chapleau, Enriching Archived Smart Card Transation Data for Transit Demand Modelling, 2008). The network density and level of service influence the transfer times (Jang, 2010) (Gordon, Koutsopoulos, Wilson, & Attanucci, 2013). In addition, transfer behaviour might also differ between user groups (Bagchi & White, 2005). Therefore, refinement of the transfer identification process has been pursued in the other two dimensions of travel behaviour.

Two types of spatial constraints are proposed:

1. *Distance covered during transfer*: the distance between the alighting stop and the next boarding stop should be low as it is assumed to be covered on foot. The distance can also be converted into walking time using a mean walking speed. The use of a buffer is recommended as walking distances depend on the infrastructure layout (Nassir, Khani, Lee, Noh, & Hickman, 2011) (Gordon, Koutsopoulos, Wilson, & Attanucci, 2013);
2. *Distance between trip ends*: in order to account for return trips after short activities, constraints can be set on the directness of the trip. This can be

done by comparing the Euclidian distance between the first boarding stop and the last alighting stop to the total distance travelled during the trip (Munizaga, Devillaine, Navarrete, & Silva, 2014) (Robinson, Narayanan, Toh, & Pereira, 2014). A similar method is to use the circuity ratio of the path travelled within a trip. If a transfer location lies outside a specified circuity area around the trip ends, it is assumed that the trip is not direct and an activity has been performed (Gordon, Koutsopoulos, Wilson, & Attanucci, 2013). In case three or more consecutive trip legs are observed that describe a tour, it is not trivial at which transfer locations an activity took place, in other words: where to cut the tour in outbound and inbound trips. The activity location can be selected using the highest likelihood based on directness of the trips (Munizaga, Devillaine, Navarrete, & Silva, 2014).

In addition, two kinds of structural constraints are proposed:

1. *No transfers on the same route*: transfers on the same route identify an activity at the transfer location. Alighting and then boarding a vehicle on the same route results in deliberate time loss, which only provides utility to the traveller in case of an activity. This holds for vehicles in both directions of the route (Munizaga & Palma, 2012) (Hofmann & O'Mahony, 2005) (Hofmann, Wilson, & White, 2009) (Chu & Chapleau, Enriching Archived Smart Card Transation Data for Transit Demand Modelling, 2008);
2. *First opportunity*: using bus route schedules it can be checked if the transfer was actually the first opportunity for the traveller. Again using the utility maximization principle, letting a opportunity for a faster transfer pass only makes sense in case of an activity (Chu & Chapleau, Enriching Archived Smart Card Transation Data for Transit Demand Modelling, 2008) (Nassir, Khani, Lee, Noh, & Hickman, 2011) (Gordon, Koutsopoulos, Wilson, & Attanucci, 2013).

Different combinations of processing rules can be used in order to identify transfers. The transfer identification procedures proposed by Chu & Chapleau (2008) Nassir et al. (2011) and Gordon et al. (2013) combine constraints in all three dimensions. They define a walking time to estimate the arrival time at the boarding stop. Using the scheduled bus times, they estimate the number of boarding opportunities between arriving at the stop and boarding. If the first opportunity is taken, a transfer is inferred. Using these constraints, Chu & Chapleau (2008) found that the automated identification of transfers in Gatineau overestimates the number of transfers with nearly 40%.

2.2.3 Identification of transfers to other systems

Transfers within the described system only constitute part of the possible transfers that can be made by travellers. The described system is limited by the available data, usually the data from one operator or region. However, the entire transport system available to travellers consists of several modes, where the PT system is a section of the total system. In addition, the PT system consists of several operators that provide their services in different areas and possibly with different modes. Service areas may be adjacent or overlapping. These constraints on the transport structure, caused by the availability of data, are not desirable in the context of trip analysis. We aim for information on the total trip, including transfers to other modes and operators.

In case the available data encompasses the data from one operator, there are three categories of transfers to other systems, which have different consequences for the construction of OD matrices:

1. *Transfers to private modes*: private modes comprise car, which can be subdivided into driver and passenger, and bicycle. Walking is considered a general component of travel and therefore is not included as mode (Nes, 2002). Transfers to private modes cannot be obtained from smart card data. However, these transfers are rare. The options to transfer between public and private modes are limited. PT users are usually bound to walking (Munizaga, Devillaine, Navarrete, & Silva, 2014) (Nassir, Khani, Lee, Noh, & Hickman, 2011) or cycling, which is popular mode of transportation in the Netherlands (Goeverden);
2. *Transfers to adjacent PT operators*: as transactions with other PT operators are not available in the data, transfers between operators cannot be obtained and result in incomplete visibility of the trip in de data. Additional trips can be made with other operators on either side of the observed trip legs. Depending on which side of the observed trip leg the transfer occurs, the origin or the destination lies outside the study area. This causes the trip to be entered in an incorrect OD cell, where the origin or destination should be an external cell. The optimal solution to the problem of transfers to other operators is combining the data from different operators, which has been done in the Netherlands by Nijënstein & Bussink (2014);
3. *Transfers to PT operators within overlapping service area*: a distinction of transfers to other operators can be made if the service areas of the operators overlap. If the traveller uses another operator to travel within the same service area, the trip is entered in an incorrect OD cell, where the origin or destination should, in this case, be an internal cell. When a different operator is used in between trip legs that are described in the available data, they can be identified by looking at the spatial-temporal distribution of consecutive trip legs (Chakirov & Erath, 2012). Depending on the used rules for transfer identification, the observed data would otherwise be interpreted as two separate trips or as one trip with a large transfer distance.

2.2.4 Conversion from stops to origins and destinations

Smart card data provide travel information on the level of used stops and stations during travels. The previous paragraphs describe the construction of matrices from these data on the level of stops. Several studies (Bagchi & White, 2005) (Pelletier, Trépanier, & Morency, 2011) acknowledged that the used stops are not equal to origins and destinations. Therefore, these matrices are not actual OD matrices, but stop-based matrices. Stop-based matrices may accurately describe the current travel demand, since the used stops are assumed to be in the vicinity of the activity locations. However, travel demand is derived demand from travellers' origins and destinations, as these are the locations where activities are performed. The actual origin and destination addresses are considered more stable indicators for future travel demand, since the used stops are, to a larger extent than origins and destinations, subject to the travel supply. The stability of indicators for travel demand is especially important in strategic level studies (Ortúzar & Willumsen, 2011).

Not all studies into the use of smart card data take this refinement into account, the stop-based matrix is then presented as OD matrix (Munizaga & Palma, 2012) (Nassir, Khani, Lee, Noh, & Hickman, 2011) (Nijënstein & Bussink, 2014). Studies that do consider the conversion from stop-based matrices to OD matrices can be divided into two categories:

1. *Clustering of stops*: several studies combine stops into stop clusters in order to reduce the number of cells in the matrix. At the lower scale this includes the aggregation of stops in opposite directions of the same line. On a higher scale the aggregation can also include the clustering of stops that serve the same catchment area (Lee, Hickman, & Tong, 2012) (Lee & Hickman, Are Transit Trips Symmetrical in Time and Space?, 2013). A second method used to cluster stops is based on movement patterns. This method includes the assessment of similarities of the used stops at both ends of the trip instead of only one trip end. Movement patterns can be clustered based on travel directions and adjacency of stops at either end of the trip (Kim, Oh, Lee, Kim, & Jung, 2014). A third method applied to cluster stops is the clustering by travel analysis zones, which equates to a direct conversion from the stops to the zone in which they are situated. This method is often used because of its applicability in transport modelling (Lianfu, Shuzhi, Yonggang, & Ziyin, 2007) (Farzin, 2008) (Zhou, Murphy, & Long, 2014) (Oort, Drost, & Brand, 2014);
2. *Allocation of origins and destinations*: trips can be allocated to specific origins and destinations. Several techniques have been applied for this allocation procedure, with different levels of resolution of the origins and destinations. Trips can be allocated to zones using a logit allocation, based on the walking distances to nearby zones and zonal characteristics (Barry, Freimer, & Slavin, 2009). In their study, Barry et al. use variable zonal characteristics, like population and employment, for different times of day, indicating the relation with the travel purpose. At a higher level of resolution, trips can also be allocated to addresses. Utsunomiya et al. (2006) use the billing address of cardholders in Chicago to assess the access distances to stops, with the assumption that the billing address corresponds to the home address. Chu & Chapleau (2010) allocate trips with student-cards to school buildings if the used stop is within 500 meter of the school address. Other trips are allocated to specific locations within the area around the stops, based on a probabilistic approach. The applied density function depends on the walking distance to the stop and the population distribution in the area. Ordóñez & Erath (2013) allocate commuting trips to work locations by minimizing the total walking distance from stops to work locations in Singapore. To do so, they use high resolution GIS data, to estimate work space capacities on parcel level and determine walking distances between stops and office buildings, and correct for the use of other transport modes used for commuting.

The clustering techniques can be perceived as OD matrices, since stops are clustered based on spatial characteristics. However, these techniques do not specifically take into account the access and egress trip legs.

Utsunomiya, et al. (2006) find that the access distance differs between rail and bus services, due to differences in the stop density and the service quality. Furthermore, access distances differ over individual stops.

2.3 Travel purpose inference

In order to take up the second issue of this study, this paragraph discusses the literature on inference of the travel purpose. As indicated by the literature in the previous paragraph, the travel purpose is closely related to the allocation of origins and destinations to trips observed in smart card data. In fact, both issues are determined by the activities performed at either end of the trip: the activity location depends on the activity type. However, the travel purpose inference is also a research

topic of its own. The purpose of travelling is a key subject of policy measures, making it an important element for grasping the influence of these measures by strategic transport models (Ortúzar & Willumsen, 2011).

The lack of information about travel purposes is a common issue for passive data collection of travel trajectories. Movements of individuals can be traced using GPS data, mobile phone data, smart card data or data from social media. All these sources provide spatial-temporal information of individuals, with different penetration rates and levels of resolution (Yue, Lan, Yeh, & Li, 2014), but lack information on travel purposes. The discovered literature on travel purpose inference focusses on two sources: smart card data and GPS data, which are discussed sequentially below.

2.3.1 Purpose inference from smart card data

Since the smart card data do not provide information on the travel purpose, but do offer detailed information on the use of PT, researchers has shifted their focus to the identification of user groups. User groups can be categorized by means of a K-means clustering method. The number of clusters can be either pre-defined (Morency, Trepanier, & Agard, 2007) or determined using a more elaborated version of clustering. Agard et al. (2006) (2009) use a Hierarchical Ascending Clustering (HAC) method to determine the number of clusters, while Ma et al. (2013) use a K-means++ clustering to find the optimal number of clusters.

The clusters are defined based on temporal variables or a combination of temporal and spatial variables. Three scales of temporal variables are defined: *times* during the day (TOD), *frequency* of travel during working days of one week and *regularity* of travel during a period of several weeks (Agard, Morency, & Trépanier, 2009) (Agard, Morency, & Trépanier, 2006). The used lines and stops can be applied as spatial clustering variables (Ma, Wu, Wang, Chen, & Liu, 2013). The user groups seem to have a high correlation with the age groups deduced from card types, which indicates the relevance for market analysis and segmentation (Agard, Morency, & Trépanier, 2009). Furthermore, these clustering methods can support short term predictions, for operators to enhance their service quality (Morency, Trepanier, & Agard, 2007) (Agard, Morency, & Trépanier, 2009) (Ma, Wu, Wang, Chen, & Liu, 2013).

Although it is argued that trip attributes from smart card data could possibly better characterize trips than the travel purpose can (Chu & Chapleau, 2010), it is also reasoned that clustering techniques cannot capture the complexity of travel patterns (Kim K. , 2014). In addition, user groups based on spatial and temporal clustering variables depend on the level of service and do not reflect a traveller's motivation for travelling, where a classification based on travel purpose does. Hence, these clusters are not the stable indicators of travel demand that are pursued in long-term planning.

Expanding the protocols of smart card data mining, the rule-based processing approach is also applied on the inference of travel purposes. Four attributes are generally considered to have explanatory value of the travel purpose:

- Activity duration
- Departure time
- Frequency
- Card type

The used rules in literature consist of constraints to these attributes:

- Activities longer than six hours are *work* activities (Chakirov & Erath, 2012);
- Activities longer than five hours with adult cards are *work* activities (Devillaine, Munizaga, & Trépanier, 2012) (Gatineau, Canada);
- Activities longer than two hours with adult cards are *work* activities (Devillaine, Munizaga, & Trépanier, 2012) (Santiago, Chile);
- First trips of the day that take place during the morning peak, and have a corresponding return trip in the evening peak, are allocated to the purpose *work* (Lee & Hickman, Trip purpose inference using automated fare collection data, 2014);
- Trips made with student cards with alighting near schools or universities are *educational* trips (Chu & Chapleau, 2010) (Lee & Hickman, Trip purpose inference using automated fare collection data, 2014);
- Trips that are the last of the day, and not the only one, are allocated to the purpose *home* (Devillaine, Munizaga, & Trépanier, 2012).

The differentiated travel purposes differ between several studies. Some studies only take into account the most prevalent travel purpose: *work* (Jun & Dongyuan, 2013) (Zhou, Murphy, & Long, 2014). If all trips are to be incorporated, the basic purposes *work*, *home* and *other* are distinguished (Chakirov & Erath, 2012), and in some papers also the purpose *education* is considered (Devillaine, Munizaga, & Trépanier, 2012) (Chu & Chapleau, 2010). Using survey data, it is also possible to consider a wider range of purposes, like shopping and business (Kusakabe & Asakura, 2014) (Lee & Hickman, Trip purpose inference using automated fare collection data, 2014). The availability of highly detailed GIS data initiates the focus on one specific travel purpose. The purpose education can be allocated to trips made with student cards when alighting takes place in the vicinity of a school address. Return trips can then be allocated to the home purpose (Chu & Chapleau, 2010). The purpose work can be allocated to trips based on parcel data of office buildings (Ordóñez & Erath, 2013).

Besides the use of rule-based processing, several other approaches have been applied, all with the aim of coupling information from survey data to smart card data. A rather direct coupling is the use of a Naïve Bayes classifier. This method assumes the same distribution of purposes relative to key variables, such as arrival time and activity duration, which are available in both survey data and smart card data (Kusakabe & Asakura, 2014). Another method is the use of a logit model for the allocation of purposes to trips. Such a model that determines the relative possibilities of a trip having a certain purpose can be estimated with survey data and subsequently applied to smart card data. Parameters with a significant influence on the distribution of chances are the activity duration, the start time and purpose-specific land-use information (Chakirov & Erath, 2012). Another method is the use of a decision tree algorithm with a learning module to classify trips into travel purpose bins (Lee & Hickman, Trip purpose inference using automated fare collection data, 2014).

Comparing a rule-based approach with a logit allocation procedure, Chakirov & Erath (2012) find that the share of correctly estimated purposes differs only slightly in favour of the logit allocation. This is mainly because the simpler rule-based model has a surprisingly high fit, with almost 87% correctly estimated purposes. Furthermore, the addition of land-use information in the logit model only marginally increases the share of correctly inferred purposes compared to a logit model without land-use information. Train (2009) issues that the share of correctly estimated choices is not a decent indicator of the model fit, since it does not capture the underlying theory of probabilities.

2.3.2 Purpose inference from GPS data

GPS sensors are commonly employed in navigation tools, smartphones and special GPS tracking devices. The sensors continuously records location, speed and direction. For a more detailed specification of GPS devices and data structures, see Wolf et al. (2001) and Stopher et al. (2008). Many studies have been performed to investigate the use of GPS data as support or replacement of travel dairy studies. The employment of GPS data generates similar issues as the use of smart card data. As GPS devices continuously record data, trip ends have to be derived in a comparable way as the identification of transfers with smart card data. In case of GPS devices in cars, the relation between exact activity location and the parking location is similar to the relation between activity locations and used PT stops (Axhausen, Schönfelder, Wolf, Oliveira, & Samaga, 2003).

A comprehensive literature review on the use of GPS data to identify person-trips is provided by Gong et al. (2014), indicating the inference of the travel purpose as one of the main research topics. The authors specify three categories of inference methods, which are also found in the literature on purpose inference of smart card data:

- *Rule-based processing*: high resolution land-use information can be combined with trip attributes such as activity duration and arrival time to infer the travel purpose (Wolf, Guensler, & Bachman, 2001) (Shen & Stopher, 2013). Purposes can also be allocated to purpose-specific points of interest (POI) in the vicinity of the trip end (Stopher, FitzGerald, & Zhang, 2008). In case of multiple POI within reach of the trip end, the closest POI can be allocated, or survey participants can be asked to provide the right purpose (Bohte & Maat, 2009);
- *Probabilistic approach*: Different POI within a specified distance of the trip end can be ranked on probabilities based on distance, the socio-demographic attributes sex, car availability and occupation status, and the trip attributes starting time, activity duration and day of the week (Axhausen, Schönfelder, Wolf, Oliveira, & Samaga, 2003). Another approach is the estimation of chances of specific purposes with a logit allocation model. Two categories of attributed are found to have a significant influence of the probabilities of purposes: the temporal attributes time of day and activity duration and the spatial attributes of land-use information. Attributes from previous trips proved insignificant (Chen, Gong, Lawson, & Bialostozky, 2010);
- *Machine learning*: based on survey data, learning models can find the relation between attributes and specific purposes and assign purposes to trips based on tree building classification. Used attributes for classification are socio-demographic attributes, trip attributes, such as activity duration and activity start time, and land-use attributes (McGowen & McNally, 2007) (Deng & Ji, 2010) (Montini, Rieser-Schüssler, Horni, & Axhausen, 2014).

Validation of the purpose inference results is possible with external surveys or by means of integrating the GPS tracks in the travel diary survey. Results of correctly classified purposes range from 43% (Bohte & Maat, 2009) to approximately 90% (Wolf, Guensler, & Bachman, 2001) (Deng & Ji, 2010). These large differences can be largely explained by the quality of land-use data, since that is the most prominent category of attributes relating to the travel purpose.

2.4 Conclusions from the literature study

This paragraph summarizes the findings from the literature study. The conclusions are categorized by findings on the discovered literature and the implications for this study.

2.4.1 Conclusions on the discovered literature

Smart card technology has entered all sorts of markets around the world with its many different applications. The data generated by smartcards are one of the main drivers of their success, as they hold valuable information on user behaviour. Likewise, the penetration in the public transport market, by means of AFC, has increased over the years and looks to be increasing still. PT Smart card data is used for many appliances: performance monitoring, market analysis and demand modelling.

Focussing on the smart card data use for travel demand modelling, the literature indicates a high potential thanks to the longitudinal character of data collection and the large sample size, which can approximate complete coverage of the described system. These factors can contribute to a significant quality improvement of current modelling techniques, since no equally rich source has previously been available for travel demand modelling.

However, smart card data cannot be seen as a direct replacement of currently used travel data, collected with surveys and counts. Compared to survey data, essential information is lacking as a result of passive data collection: travel purpose, origins and destinations are unobserved. This information is required for the interpretation of the data. The travel purpose, or activity, provides the reason for travelling, which is essential to long term demand forecasts that deal with policy changes. Origin and destination provide the locations of activities, which are similarly essential to long term demand forecasts as they are subject to changes in land use. Furthermore, the interpretation of check-in and check-out transactions in terms of trips, tours and transfers, depends on definitions used in rule-based processing.

The method of OD matrix construction with smart card data depends on the data structure and the utilisation of the OD matrix. Regarding the data structure, various datasets exist from different AFC systems. Essential elements are the availability of the alighting transactions and the integration with AVL systems. Quality differences exist in the rule-based processing procedures to derive transfers within the system. Additional processing rules can be added to basic rules in order to allow for more exceptions. In contrast, transfers to other systems are not a subject of interest in the discovered literature. Regarding the utilization of the matrices, most studies focus on short term planning and therefore do not incorporate the conversion from matrices from stop-level to zone-level. For long term planning, however, stops do not have the desired stability of indicators for trip production and attraction.

The travel purpose inference is acknowledged as a main issue for strategic planning as it incorporates the motivation for travelling, which is subject to policy measures and societal changes. Therefore, it is a recurring topic in literature on the implementation of new data sources in travel demand modelling. All passive data collection methods encounter this problem and, despite differences in data structure, similar methods of purpose inference have been studied.

Inspired by the large amount of data, several studies deal with the absence of travel purpose information by application of user group clustering. For short term planning,

this method provides insight in travellers affected by disruptions and rerouting. Yet, this method does not suit long term planning as it does not determine the motivation to travel.

The most simple method builds on the data mining procedures and applies rule-based processing to infer travel purposes. These rules involve crude simplifications and cannot differentiate between many travel purposes. More sophisticated methods are based on probabilities and machine learning principles. These methods identify the same variables to be explaining information on the travel purpose. These variables can be categorized in three groups: (1) trip characteristics like activity duration, start time, frequency and card type, (2) socio-demographic characteristics like age and occupancy status and (3) land-use characteristics. The influence of land-use characteristics highly depends on the resolution of the available data. The resolution may vary from aggregated zonal data to high resolution land-use data, which contain individual buildings.

2.4.2 Implications for this study

Assessing the discovered literature, we can conclude that the research questions of this study have been studied frequently in similar contexts. Against the background of this literature, however, we have also found that dissimilarities in available data and eventual employment have great impact on the used methods to answer them.

Regarding the data processing rules to interpret the crude data, the literature provides several options, with different levels of detail, for the identification of transfers and activities. Several issues arise with the interrelation between them. First, inconsistent use of PT results in single trips, for which it is not possible to derive a tour. Another possible result from inconsistent use is the occurrence of "gaps" between subsequent alighting and boarding. These gaps can be interpreted as an activity or as an unobserved transfer: a trip leg made in another system, which could be a private mode or another operator. Second, transfers on the same line distinguish activities, but not all of these activities might be relevant for this study. Different processing rules result in contradictory "observations". The implications of these issues are especially relevant for modelling context based on tours.

Rule-based processing can also provide feasible solutions for the purpose inference, although this reduces the possibilities to incorporate a larger scale of travel purposes. Other feasible methodologies for the purpose inference have a wide range of detail. The Naïve Bayes Classifier used by Kusakabe & Asakura (2014) seems suitable for the enrichment with survey data, but depends on the stability of this data. Since the survey available for this study has been terminated, this method might not be very stable. More deliberate approaches by (Chakirov & Erath, 2012) and (Lee & Hickman, Trip purpose inference using automated fare collection data, 2014) also seem appropriate for this study. Although only applied to high resolution land registration data, the literature has also indicated that the combined estimation of the exact location and the activity holds potential. Therefore the option of combining the models for zone allocation and purpose inference might be an fertile approach.

2.4.3 Overview of applicable methodologies to answer the research questions

The literature has provided several methodologies to tackle the two main problems at hand. Table 2 presents the methodologies that are deemed applicable for this study, with the corresponding attributes and data sources used.

Table 2: Overview of applicable methodologies found in literature

Problem	*Approach*	*Applied attributes*	*Applied data sources*
Zonal allocation	Direct conversion to zones	- Stop locations	- Stop coordinates - zonal boundaries
	Clustering of stops	- catchment areas - movement patterns	- OD matrix - zonal centroids
	Logit allocation	- distances between stops and zones	- stop coordinates - land-use data
Purpose inference	Rule-based processing	- activity duration - departure times - frequencies - card types - purpose specific land use attributes	- survey data - land-use data
	Naïve Bayes classifier	- arrival time - activity duration	- survey data
	Logit allocation	- activity duration - departure time - purpose-specific land use attributes	- survey data - land-use data
	Machine learning three building classification	- Gender - age - job status - activity duration - departure times - purpose-specific land use attributes	- survey data - land-use data - traveller data

The literature does not provide options for the identification of concession traversing transfers. For this specific issue, a straightforward approach has been pursued, based on the available data.

2.4.4 Overview of assessment criteria for the methodology choice

The methodologies found in literature have indicated differences in the eventual application of the results. The eventual application determines the desired quality of the method. On the other hand, the feasibility of a methodology depends on the available data and budget constraints. Table 3 contains the seven assessment criteria that were used in the methodology choice.

Table 3: Methodology assessment criteria

Quality of the method	• *Feasibility of the method*
Behavioural richness	Fit for available data
Level of detail	Budget constraints
Durability of the method	Flexibility in development
Interpretability	

The methodology choice was based on these criteria. With knowledge of the available data sources, and their qualities and limitations, a selection of feasible methods was derived for both the zonal allocation and the purpose inference. Subsequently, these methods were assessed on their qualities, in relation to the eventual application of the OD matrices in transport models. In addition, feedback on these criteria provided input for the evaluation, which in turn led to recommendations for future research.

The next chapter focusses on this process by first describing the available data sources and, subsequently, the applied methodology.

3 Methodology and data

This chapter considers the methodology to answer the research questions and the employed data sources. In the introduction, we introduced the limitations of smart card data and the aspired coupling with survey data, in order to construct purpose specific OD matrices. Subsequently, the literature review provided an overview of possible methods to achieve this objective and criteria to assess them. In this chapter, we describe the applied methodology and the motivation for choosing it over the other alternatives.

In order to do so, the chapter starts with the specification of the features of the data sources, starting with the OV-chipkaart (paragraph 3.1), followed by the available travel surveys (paragraph 3.2) and the land use data (paragraph 3.3). Based on the available sources and the methodology assessment criteria derived from literature, we motivate the choice of methodology for the enrichment models (paragraph 3.4). Subsequently, the research outline is presented (paragraph 3.5), followed by a more detailed description of the three research phases (paragraphs 3.6 to 3.8). Finally, conclusions about the available data sources and the applied methodology are listed (paragraph 3.9).

3.1 The Dutch smart card: the OV-chipkaart

The Dutch smart card, the OV-chipkaart, is employed by all public transport operators in The Netherlands. Almost all travel products have been converted to the smart card. Therefore, the OV-chipkaart approximates complete coverage of the public transport system, recording all trips made by train and by bus, tram and metro (BTM). The exceptions consist of several operators that still sell un-chipped tickets at the driver and trips with missing check-in or check-out transactions.

The literature on the use of smart card data shows substantial differences between smart card systems applied for revenue collection in public transport. Similar to the Singapore EZ link card, the OV-chipkaart applies revenue collection with distance-based fares. Hence, both at boarding and alighting a transaction is required to determine the distance travelled. The smart card system is coupled with a GPS tracking system, which directly translates the location of transactions to stops. The accurate registration of transaction times allows operators to differentiate their fares between peak and off-peak periods. Furthermore, the personal OV-chipkaart allows for fare differentiation based on age groups, children and seniors, by registering the birth year. The anonymous OV-chipkaart does not register a birth year.

The OV-chipkaart registers the data per transaction. Table 4 contains the attributes relevant for this study.

Table 4: Available attributes in OV-chipkaart data

Attributes	*Specifics*
Card number	hashed, but uniquely identifiable
Transaction sequence number	Counts transactions by card
Transaction type	check-in or check-out
Date and time	accurate to seconds

Stop/station	coded
Entry stop/station	only for check-out transactions
Concession	also indicates operator
Mode	Train, bus, tram or metro
Line	only for bus and tram
Distance travelled	based on route
Card type	Personal or anonymous
Travel product	Contract type or e-purse
Fare	Full, reduction or unlimited travel

From these attributes, additional information can be derived. Transactions can be aggregated into trip legs via the transaction sequence number, transaction type and the entry station. Subsequently, trips can be aggregated into trips based on a combination of transaction times and locations, depending on the distinction between transfers and activities. With a similar procedure, but different processing rules, trips can be aggregated into tours. The metadata of the available OV-chipkaart data and the applied processing-rules are described in paragraph 4.2.

3.2 The WROOV surveys

The survey data employed for this study comprise the WROOV-light studies. These studies consisted of a yearly travel survey, running from 2003 up to 2009, with the purpose of revenue allocation of the NVB ticketing system. The NVB contained all national tickets for bus and light rail, but did not include the student card and regional tickets. Moreover, the NVB tickets were not valid on the majority of national railways, hence the WROOV survey does not include train travels.

During one month a year, purchased travel products were accompanied by a survey form, requesting the travels made with the product. For *strippenkaart* tickets, one trip was to be declared. For contracts, the most frequent trip was asked, with the corresponding frequency, and one occasional trip. For all trips, the additional question was asked if that trip was also made in the opposite direction. This resulted in 110.000 to 150.000 completed surveys a year, adding up to a total of 1.7 million trips over seven years. This sample size is extraordinary large for a travel survey, especially one focussed on bus and light rail.

The WROOV data contains information on the traveller and on the travels made. However, the dataset is not structured based on trips, but on revenue allocation elements. These can subsequently be aggregated into trip legs, trips and tours, based on the available attributes. After aggregation into trips, the dataset contains the attributes presented in Table 5.

Table 5: Available attributes in WROOV data

Attributes	*Specifics*
Survey form number	hashed, but uniquely identifiable
Gender	
Age	at the time of the survey
Tour number	always 1 for *strippenkaart* tickets,

	max 2 for contracts
Trip number	away trip or return trip
number of trip legs in trip	based on indicated transfers
Origin	at the level of PC6
Destination	at the level of PC6
Boarding stop	coded
Alighting stop	coded
Weekday	Monday to Sunday
Departure time	Accurate to minutes
Mode	Bus, tram or metro
Route	for bus, tram and metro
Distance travelled	mostly based on route
Travel product	Contract type or *strippenkaart*-tickets
Fare	Full, reduction or unlimited travel
Concession	also indicates operator
Frequency	number of tours per week
Travel purpose	7 distinct purposes, as well as *multiple* and *other*

The WROOV data does not require data processing to determine transfers or activities, as these are indicated by the respondent. The activity duration can be derived based on the departure times of consecutive trips. Since the alighting time is lacking, the activity duration includes the travel time of the away trip. Therefore, the activity duration derived from WROOV data is slightly overestimated.

One alternative option for survey data in the Amsterdam region was the MON/OViN survey. This survey consists of a yearly survey that is still running, although a change in the data collection method resulted in a trend reversal between 2009 and 2010. However, the complete sample size is much smaller than WROOV, with the essential distinction that MON/OViN includes all transport modes. Hence, the sample size per year for public transport is less than 50 times smaller compared to WROOV (Kuhlman, 2014). Moreover, the MON/OViN data does not include the used public transport stops, which are essential for the allocation of origin and destination zones. Consequently, the MON/OViN data is considered not suitable for this study and has not been employed.

3.3 Land use data

Besides the previously introduced travel data sources, this study has also employed land use data. Other studies have indicated that land use data are valuable for both the zonal allocation as well as the purpose inference. The land use data available for this study originates from the Amsterdam Region transport model VENOM, which is introduced in paragraph 4.1. The land use data contains averaged data per zone in the VENOM zonal grid. Table 6 presents the attributes in the land use data relevant for this study. We classified the attributes, based on their expected influence on the zonal allocation.

Table 6: Available attributes in land use data

Attributes	*Classification*	*Specifics*
centroid coordinates	Geographical	based on gravitational centre of the zone
area	Geographical	rounded to hectares
residents	Home-end	
students	Home-end	
working population	Home-end	
cars	Home-end	registered by residents
households	Home-end / Activity-end	
jobs	Activity-end	
student places	Activity-end	for 5 school categories

The zonal data contains attributes related to the activity end of the purposes work and education, but does not include any data specifically related to the purpose shopping.

3.4 Qualitative comparison of the OV-chipkaart and WROOV

A qualitative comparison between the OV-chipkaart and WROOV has been performed in order to determine the possibilities and appropriateness of transferring information between these sources. First, key variables have been identified by comparing the available information in the sources. Second, the coverage and target population of both sources have been compared, which has provided input for the construction of comparable data sets.

3.4.1 Key variables for coupling of information

By comparing the available attributes in both travel data sources, key variables can be determined. These variables are the instruments of transferring information between sources. In order to do so, the information to be added is to be expressed in terms of these key variables.

Table 7: Key variables for transferring information

Key variables between OV-chipkaart and WROOV data
Activity duration
Frequency
Departure time
Travel distance
Contract duration
Fare
Mode
Operator
number of legs within trip
number of trips within tour
Zonal data at the stop locations

3.4.2 Sample sizes, target populations and data collection periods

In order to determine their validity as medium of transfer, the description of key variables in both data sets has been in compared. The more equally distributed the values of key variables are, the higher their validity. The results of this quantitative comparison between data sources are presented in paragraph 5.4. When perceived as valid, attributes have been included in the model estimation process, where their explanatory value has been assessed.

Table 8: Qualitative comparison of the data sources OV-chipkaart and WROOV

	WROOV	*OV-chipkaart*
Information	Surveyed trips Used modes Used stops Departure time Origin and destination Travel purpose	Observed trips Used modes Used stops Boarding and alighting time
Period	yearly from 2003-2009	1 week in 2014
Coverage	National tickets No local cards No Student cards	All tickets
Location	Amsterdam + Waterland	Amsterdam + Waterland

3.5 Methodology of enriching smart card data

In order to construct purpose-specific OD matrices with OV-chipkaart data, three pieces of information have to be coupled: the origin, the destination and the travel purpose. The WROOV data contains all these three elements, and therefore these can be applied as source of the required information. The literature study has provided several options to allocate trips to zones and infer the travel purpose based on survey data (see paragraph 2.4.3). In addition, a set of assessment criteria was constructed based on the evaluation of this study in context of reference studies (see paragraph 2.4.4).

The number of feasible methods is limited by the fundamental objective and the specifications of the available data. The objective is to create matrices that are applicable as base matrices in transport models. Therefore, the allocation of origins and destinations needs to comply with the zonal structure of the model. Furthermore, the level of resolution of the zonal allocation based on land-use data is limited to an aggregated level due to data constraints. Since origins and destinations recorded in the WROOV data are not recorded as addresses and the available land use data are aggregated at the zonal level, the allocation of origins and destinations to specific addresses is not feasible.

3.5.1 Zonal allocation

Initially, three applicable methods for the allocation of origin and destinations were derived from the literature:

1. Direct conversion of stop locations to zones;
2. Clustering of stops;
3. Logit allocation models.

The direct conversion of stop locations to zones inevitably complies with the zonal level of resolution. However, the accuracy of direct conversion depends on the size of the zones in relation to the access and egress distances. Within a high resolution zonal grid, the chance increases that origins and destinations are not situated in the same zone as the stop. The clustering of stops results in origins and destinations which do not comply with the zonal structure of a transport model. Therefore this method would not result in OD matrices with the desired level of resolution.

Hence, the logit allocation models are perceived as the most appropriate method for this specific combination of research objectives and available sources. Consequently, this study has applied logit allocation models for the allocation of origins and destinations to trips. These origins and destinations consist of the distinct zones classified in the transport model, with the corresponding land use data. Logit models estimate the chance of a specific option being "chosen", relative to the other options. The model parameters to be estimated determine the influence of specific land use variables on the chance a trip originates from or terminates in a specific zone.

In order to investigate the level of resolution at which logit allocation models actually perform better than a direct conversion of stop location to zones, this method has also been applied as a reference. Paragraph 3.8 continues in more detail about the framework of the enrichment models.

3.5.2 Purpose inference

Regarding the purpose inference methods, the literature has provided four applicable methods:

1. Rule-based processing;
2. Naïve Bayes classifier;
3. Logit allocation;
4. Machine learning classification tree.

For this study, we aimed for a disaggregate approach, to match the disaggregate nature of the OV-chipkaart data, eliminating the Naïve Bayes Classifier as desired approach. The rule-based processing approach applies crude simplifications, with resulting errors that might cancel out at an aggregated level. The probabilistic approach of logit allocation fits better to the disaggregate OV-chipkaart data and the uncertainties in the distribution of travel purposes.

The machine learning approach was also considered a feasible method, but for practical reasons we applied the logit allocation. Since logit allocation is also applied for the zonal allocation, both model types are estimated in the same software package Biogeme, which is freely available. Using the same modelling framework and software for both problems limited the required time to get familiar with the software.

Moreover, a combination of both problems in a single model was considered. A combined model, estimating both the destination zone and the travel purpose, proved to be possible. However, the large number of attributes and alternatives[5] made this model hard and time expensive to interpret. Therefore, this approach has not been sustained.

Similar to the zonal allocation models, the purpose inference models have been estimated in both the trip-based and the tour-based approach. In addition, a simple rule-based processing approach has been applied. Consequently, purpose-specific OD matrices have been constructed for all three approaches. These allowed for a comparison between the approaches, and thereby drawing conclusions on the differences.

3.6 Research outline

Following the literature study and the methodology choice, the remainder of the research has been set-up in three different phases. Primarily, the data analysis was performed in order to provide insight in the correlation in the WROOV data between key variables and the lacking information. In addition, a quantitative comparison on key variables between WROOV data and the OV-chipkaart data provided insight in the transferability of information and the appropriateness of attributes as explanatory variables in the enrichment models.

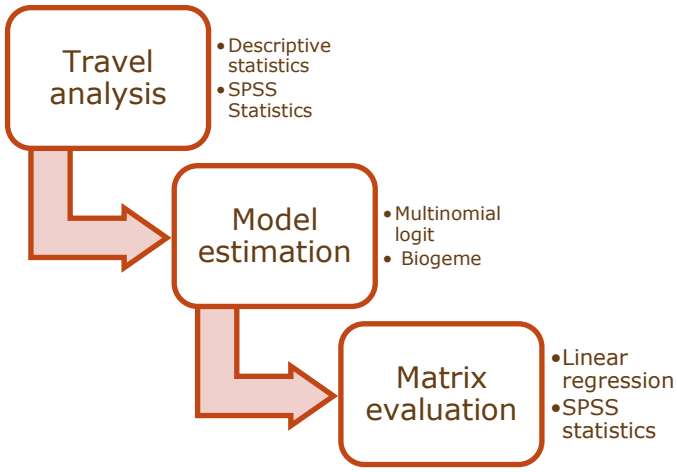


Figure 4: Research set-up

Successively, in the second phase, the estimation of the enrichment models has been performed. This process consisted of the selection of explanatory variables and calibration of the model parameters. Where the initial selection of explanatory variables was based on their explanatory value described in literature and their correlation in the data, this selection process was based on the model performance and the interpretation of the model parameters. Two approaches have been applied, starting with the less demanding trip-based approach, which was successively enhanced to a tour-based approach. For both approaches, this resulted in models that allocate origins and destinations to trips observed in OV-chipkaart data and models that infer the travel purpose of these trips.

[5] For the combined zonal allocation and purpose inference model, the number of alternatives is equal to the number of available zones, multiplied with the number of distinct purposes.

Finally, in the third phase, the estimated models were applied to both WROOV data, and to OV-chipkaart, in order to construct purpose specific OD-matrices for both sources. The resulting matrices were compared and resulted in the evaluation of five different aspects:

1. Model validation: the application of models onto WROOV data served the model validation by comparing the resulting matrices with the observed matrix;
2. Source comparison: the OD matrices resulting from the same modelling approach on both sources allowed for a comparison between the movement patterns described by WROOV and by the OV-chipkaart;
3. Model approach comparison: the comparison on different levels of resolution allows for the assessment of the added value of more complex model approaches in relation to the level of resolution;
4. Face validation OV-chipkaart matrices: the OD matrices based on OV-chipkaart data allow for a face-validity check based on several high-profile movement patterns.

The research outline is visualized in Figure 5. In the following paragraphs, the methods applied in the successive research phases are described in more detail.

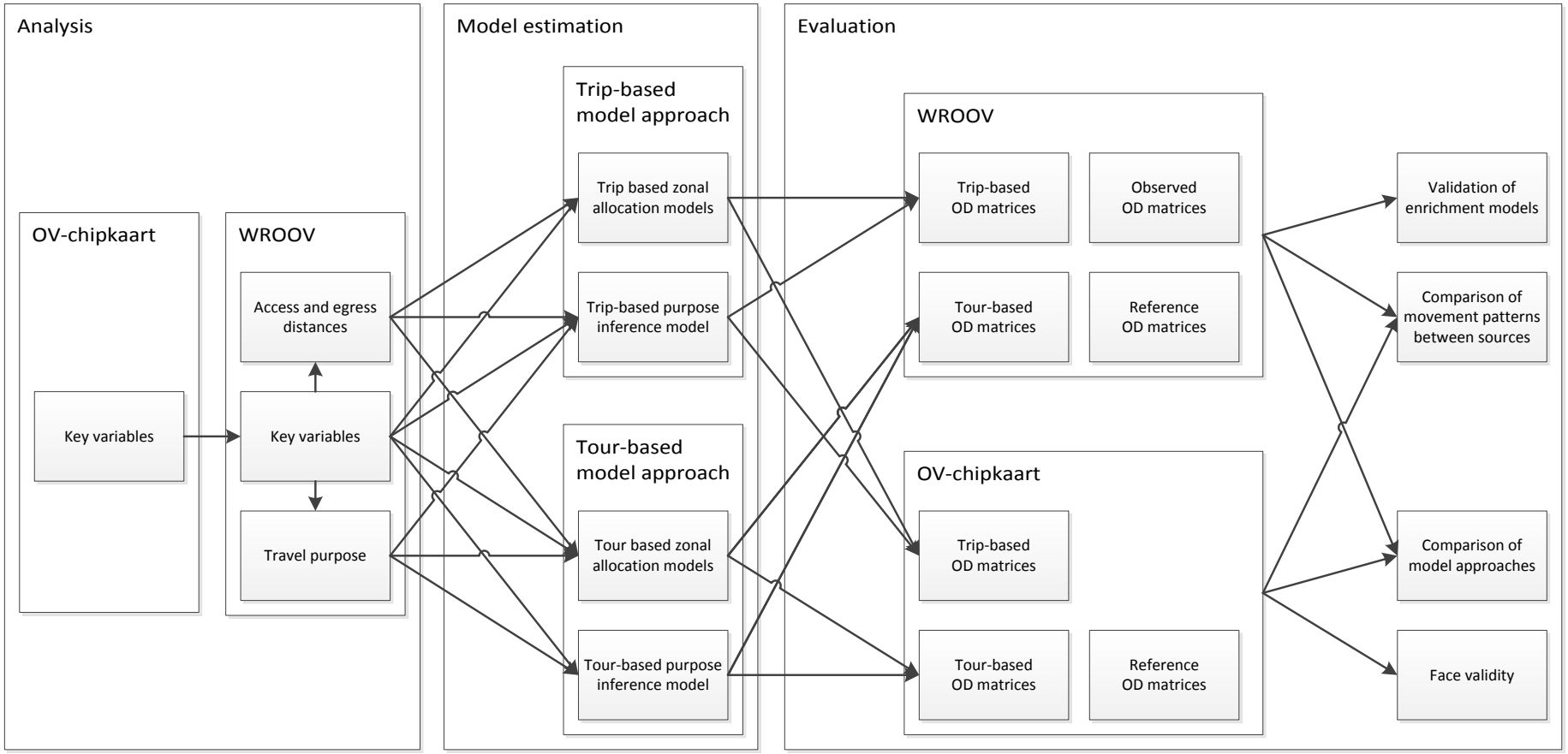


Figure 5: Research outline

3.7 Analysis framework

The literature review has indicated attributes which might have explanatory value in the estimation of access and egress distances as well as attributes correlated with the travel purpose. Several of these attributes are available in both OV-chipkaart data as well as WROOV data, and therefore belong to the key variables which can be used to transfer information between these sources.

The analysis of WROOV data was focussed at relating the access and egress distances and the travel purpose to key variables. This has been done by analysing the descriptive statistics of these variables in the software package of IBM SPSS Statistics and visualization of the distributions over key variables. Regarding the access and egress distances, these analyses have been performed for both the trip-based definitions of origin and destination, as well as the tour-based definitions of home-end and activity-end, in order to evaluate the possible explanatory value of both approaches.

Successively, the description of key variables within both datasets has been compared in order to determine the appropriateness of key variables as medium of the information transfer. This analysis is based on relative frequencies of key variables in both sources.

Key variables that were perceived as appropriate or valuable to the estimation of lacking information were selected into the initial set of attributes for the estimation of the enrichment models. The results from the travel data analysis are presented in chapter 5.

3.8 Modelling estimation framework

The framework of the enrichment models for both the zonal allocation and the purpose inference consisted of logit models, based on their employment in discrete choice modelling. Discrete choice models have been applied in many feeds, for example the travel demand related mode choice and route choice problems. Here, we describe the basic workings of this modelling framework. For more deliberate explanations and background, we advise reading the work by Ben-Akiva and Lerman (1985) and Train (2009).

3.8.1 Theoretical background

Discrete choice modelling is based on random utility theory[6]. This theory states that individuals optimize their utility in the choices they make. The utility function of alternatives can be divided into a measurable part, which can be explained by attributes, and an error term.

$$U_{ni} = V_{ni} + \varepsilon_{ni}, \qquad \forall i \in I \qquad 3.1$$

With:
- U_{ni} = utility of alternative *i* for individual *n*
- V_{ni} = systematic utility of alternative *i* for individual *n*
- ε_{ni} = error term of alternative *i* for individual *n*
- I = choice set

[6] An alternative to the optimization of utility is the minimization of regret. This theory has been developed based on the idea that not all choice processes are based on utility, but different choice strategies exist. (Chorus, Arentze, & Timmermans, 2008)

The measurable part of the utility function consists of a vector of attributes multiplied with their corresponding utility coefficients. These coefficients are assumed to be constant over individuals.

$$V_{ni} = \sum_{k=1}^{k} \beta_k * X_{ik}, \qquad \forall i \in I \tag{3.2}$$

With: V_{ni} = systematic utility of alternative *i* for individual *n*
β_k = Utility coefficient for attribute *k*
X_{ik} = value of attribute *k*

The probability of an alternative being chosen depends on the utilities of all available alternatives. Depending on the assumptions regarding the error term, different model specifications result in different expressions of the probability formula. Multinomial logit (MNL), which is the most simplified and commonly used form of a logit model, assumes the error term to be Gumbel distributed with the Independent of Irrelevant Alternatives (IIA) property. This implies that the choice set contains all relevant alternatives and error terms are uncorrelated between individuals and alternatives. While this may not be realistic in many situations (Train, 2009), this causes the error terms to cancel out in the probability formula, leading to the MNL formula in equation 3.3 (Ben-Akiva & Lerman, 1985).

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_{j \in I} e^{V_{nj}}} \tag{3.3}$$

With: P_{ni} = the probability of individual *n* choosing alternative *i* from choice set *I*
V_{ni} = measurable utility of alternative *i* for individual *n*

The model estimation process consists of the optimization of utility coefficients of the specific attributes to fit a dataset with choice observations. Several methods are available for this process. We have used the software package Biogeme, which uses optimization of the log-likelihood of the model describing the data. In addition to the assessment of the model fit, the significance of individual parameters can be tested by means of the t-test. Based on this test, relevant attributes to be included in the model can be selected.

3.8.2 Application in allocation problems

The problems at hand do not comply with the theoretic definition of a choice problem. Although the home zone of a traveller is related to the residential location choice, this is not a choice we consider here. On the activity end of the trip, the traveller generally has made the choice of destination long before alighting at the PT stop. The travel purpose is also a choice that is made before the trip is initiated. Hence, the model framework is in this case only used for the optimization of the probability that a trip originates or ends in a specific zone, or is made for a specific purpose. Hence, the discrete choice modelling framework is used for three separate allocation problems.

3.8.3 Distinct zonal allocation models

Two approaches have been pursued in the estimation of logit allocation models: a trip-based approach and a tour-based approach. In the-trip-based approach, trips are handled as uncorrelated units of travel. This resulted in two distinct models for both trip ends: one for the origin zone allocation and one for the destination zone

allocation. Conversely, in the tour based approach, trips are treated as correlated within tours. This ensures consistency of origins and destinations between trips within the same tour. The tour-based approach resulted in four distinct models. Besides the home zone and activity zone allocation models, which are estimated for tours, these also include origin zone and destination zone models for non-home-based trips (Figure 6).

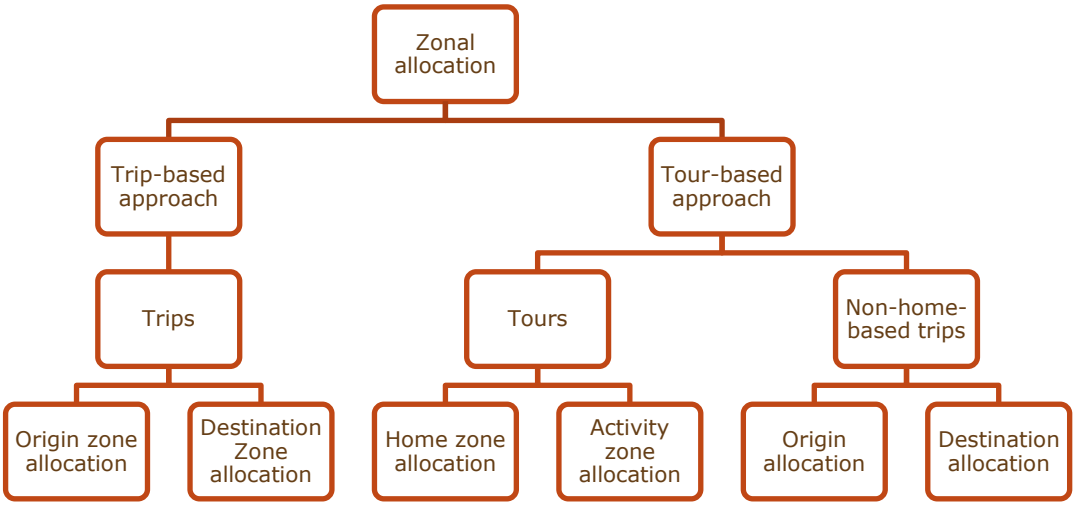


Figure 6: Classification of zonal allocation models

3.8.4 *Generic model structure for zonal allocation models*

The choice set generation yields the available alternatives per stop. Yet, the zonal allocation needs to be applicable for every stop, since it was not preferable to estimate a distinct allocation model for every stop. Stops at different locations have different zones as alternative origins or destinations. Consequently, the available alternatives have to be generic.

Since the utility of alternatives does not depend on trip characteristics, as these are equal for all alternatives, the utility only depends on zonal characteristics. In order to create generic alternatives, the alternative zones are numbered based on their share of the catchment area and successively matched to the zonal data obtained from the VENOM model (see Figure 7). The data file then consists of records that each consist of the observed trip-end zone and the zonal data for every alternative. If the number of available alternatives is lower than the maximum number of alternatives, the remaining alternative numbers are marked as unavailable with an availability identifier. In order to prevent the ranking of alternatives to influence the choice probabilities, the alternatives are randomly distributed over the maximum number of alternatives. By including the zone number and availability identifier in the randomization, it is still possible to identify to which zone each alternative corresponds.

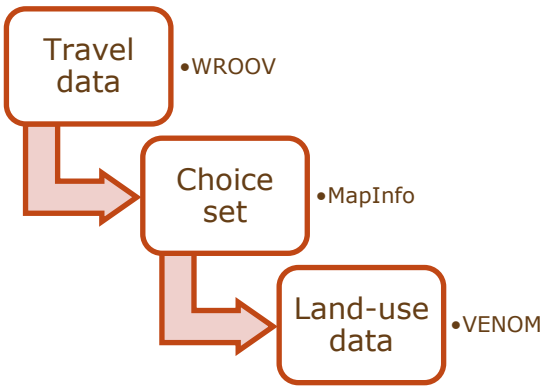


Figure 7: Data handling process of generic model structure for zonal allocation

Table 9 provides a simplified version of the specification of choice alternatives in the Biogeme model files, indicating the generic character of the alternatives. The number of specified alternatives (n) depends on the choice set generation. To which zone the alternatives correspond depends on the used stop and can be determined from the data file. The zonal characteristics, attributes X_1 to X_m, are included for all n alternatives, but set to the Biogeme missing value if the alternative is unavailable.

Table 9: Generic model specification for zonal allocation in Biogeme

Alternative	*Availability*	*Utility specification*
$Zone_1$	$Availability_1$	$B_1 * X_1(Zone_1) + B_2 * X_2(Zone_1) + .. + B_m * X_m(Zone_1)$
$Zone_2$	$Availability_2$	$B_1 * X_1(Zone_2) + B_2 * X_2(Zone_2) + .. + B_m * X_m(Zone_2)$
⋮		
$Zone_n$	$Availability_n$	$B_1 * X_1(Zone_n) + B_2 * X_2(Zone_n) + .. + B_m * X_m(Zone_n)$

The generic specification of alternatives provided the opportunity to apply identical model specification files for both the trip-based as well as the tour-based approach. Moreover, adapting the number of alternatives ensured the same structure was also applicable for the alternative approaches of choice set generation.

3.8.5 Distinct purpose inference models

Similar to the zonal allocation models, the purpose inference models have been estimated with the trip-based approach and the tour-based approach. This ensures that the complete construction process of purpose-specific OD matrices is executed within these two approaches.

Consequently, three distinct purpose inference models have been estimated: purpose inference model for all trips, a purpose inference model for tours, and a purpose inference model for non-home-based trips.

3.8.6 Model structure of purpose inference models

In contrast to the zonal allocation models, the purpose inference models do not require a generic selection of alternatives. The available alternatives consist of a predetermined number of distinct purposes. The WROOV surveys differentiated nine travel purposes, including *multiple* and *other*. Previous research indicated that many of these purposes cannot be distinguished based on travel patterns (Kuhlman, 2014). Based on their frequency, four distinct purposes are identified as the most relevant purposes to be included in the model:

- Work;
- Education;
- Shopping;
- Other.

The attributes of the purpose inference model consist of trip characteristics. Since information can only be transferred via attributes that are available in OV-chipkaart data as well, these characteristics are limited to key variables.

In addition to these trip characteristics, land-use data at both ends of the trip has been implemented as attributes. Since the origins and destinations are not exactly at the used stops, the land-use data has been averaged over an area within a radius of 400 metres around the stop. This distance is chosen conservatively, since the averaging effect of the land-use data relates to the area, which increases quadratic with the radius. As a result of the unknown origin or destination, the land-use data values applied might not be consistent with the real land-use data at the origin or destination.

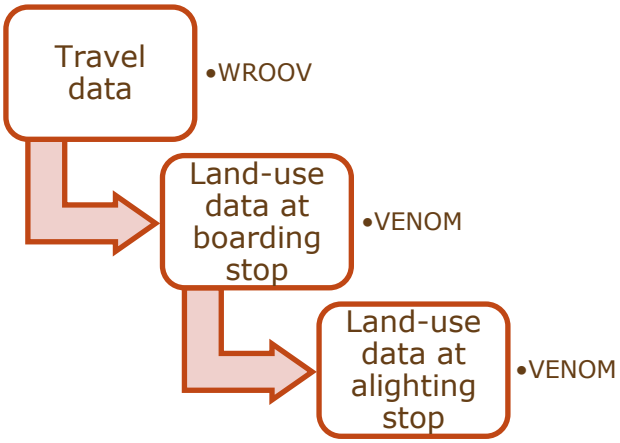


Figure 8: Data handling process of purpose inference models

The results from the estimation of the enrichment models are presented in chapter 6.

3.9 Matrix evaluation framework

The model estimation phase produced three different sets of enrichment models: the rule-based reference models, the trip-based models and the tour-based models. All three sets have been applied to both the WROOV data and the OV-chipkaart data, resulting in three different sets of OD matrices per data source. In addition, we also have observed OD matrices from the WROOV data, leading to a total of seven sets of OD matrices. All of these sets consist of OD matrices for three different divisions in space, four divisions in time, and five divisions by purpose. Hence, a total of 7 * 3 * 4 * 5 = 420 specific matrices are available for the evaluation. Not all of these matrices are worthwhile to compare with each other.

First, in order to draw conclusions on the model validity, the OD matrices based on observations in WROOV have been compared with the OD matrices constructed with the three modelling approaches. Since observed OD matrices from OV-chipkaart data are not available, we assess distinct OD pairs on face validity. Second, in order to draw conclusions on the comparability of the sources, the OD matrices of WROOV and

OV-chipkaart based on the tour-based models have been compared. Finally, in order to draw conclusions on the differences between sources, the OD matrices of OV-chipkaart based on the tour-based models have been compared to the OD matrices based on the trip-based models and the reference models. An overview of the exact matrix comparisons is presented in chapter 7.

Table 10: Matrix divisions

Matrix division dimension	*Divisions*
Space	PC3 PC4 VENOM zones
Time	Average working day Morning peak/Off-peak/Evening peak
Purpose	All purposes Work/education/shopping/other

The method of evaluation of OD matrices is not straightforward. The number of cells within OD matrices increases quadratic with the number of distinct zones in the study area of the transport model. With a high level of resolution, this implies a large number of specific movement patterns between zones. Moreover, the numbering of zones within a study area, usually based on postal codes, might not be related to their geographical situation. Therefore, comparing OD matrices is limited to techniques which have their limitations (Allos, Merrall, Smithies, & Fishburn, 2014) (Pollard, Taylor, & van Vuren, 2013).

The British Department for Transport applies a quite rigid method by comparing OD matrices based on a linear regression of the cell values. This method is applied to determine the difference between base matrices before and after the calibration process (Department for Transport, 2014). This objective is similar to the comparison pursued in this study.

Another assessment option is the Mean Structural Similarity (MSSIM) index. This method compares cells in the OD matrix per block in order to assess structural differences (Djukic, Hoogendoorn, & Lint, 2013). However, techniques based on structural similarity are still under discussion (Pollard, Taylor, & van Vuren, 2013). Since the structural similarity in the matrices based on OV-chipkaart data is guaranteed based on stop locations, this method has limited additional value in the evaluation of the applied estimation models.

In addition to the evaluation with linear regression, the comparison of different modelling approaches based on the network assignment would provide additional assessment options which directly relate to the influence of the differences between modelling approaches on the network performance. However, this comparison would require many time-consuming model runs and insight in the assignment models, which is considered a very interesting follow-up study.

Hence, the applied method of evaluation consists of linear regression. The equation of the linear regression line is presented in 3.4. This technique estimates the cell values of one OD matrix (the dependent variable) based on the cell values of another OD matrix (the one independent variable). This results in three different parameters that can be assessed:

1. The r^2 statistic: this statistic is a measure of the model fit, which represents the explained variance in the dependent variable by the independent variables. Its value ranges from 0, indicating no explained variance at all, to 1, indicating fully explained variance;
2. The *a* parameter: this parameter represents the intersection of the regression line with the y axis in the regression formula;
3. The b parameter: this parameter represents the slope of the regression line, which is the derivative of y in respect to x.

$$y = a + b * x \qquad 3.5$$

In case of complete equal matrices, the results would indicate a r^2 statistic of 1, an parameter a with value 0 and a parameter b with value 1. The closer the statistics approach these values, the more matrices are alike. In order to assure enough comparability between matrices before and after calibration, the Department For Transport requires r^2 values of 0.99. This value is not considered feasible for comparing OD matrices based on different modelling approaches.

The results from the matrix evaluation are presented in chapter 7.

3.10 Conclusions regarding the methodology and data

In order to provide an overview of the implications of the available data sources and the applied methodology on the results of this study, this paragraph lists the essential conclusions from this chapter.

3.10.1 Conclusions regarding the data sources

The literature already indicated the importance of the data structure of smart card data in relation to the quality of its employment in travel demand studies. In addition, we found that this also holds for the use of survey data and land-use data used to enrich smart card data.

The Dutch OV-chipkaart has a rich data structure, since it contains both boarding and alighting transactions, as well as the used travel product. Moreover, the transactions are automatically coupled to stops by an integrated GPS system. Besides the rich data structure, the coverage of the system is also high compared to similar systems around the world, as it is the only valid ticketing system in most regions.

The WROOV data contain all the required information for the enrichment of smart card data. This survey includes the used stops, which allows for an analysis of the access and egress distances, as well as the travel purpose. The sample size is relatively large for a travel survey and fully focussed on transport with bus and light rail. The only drawback of this source is that the survey has been terminated in 2009, which limits the appropriateness of the results over time.

The land-use data from the VENOM model match the zonal grid, which allows for a straightforward linkage with zones as alternatives in the logit zonal allocation models. The data contain specific attributes related to the home-end and to the activity-end for purposes work and education. However, no directly related attributes are available for the activity-end of the purpose shopping.

3.10.2 *Conclusions regarding the methodology*

The methodology of enrichment of OV-chipkaart data is based on logit allocation. These models estimate the probability a specific alternative is "chosen" relative to the other alternatives, based on their explained utility. During the model estimation, the influence of attributes on the utility of an alternative is determined. Concerning the zonal allocation models, these alternatives consist of zones nearby the used stop. Attributes that influence the utility of these attributes are zonal characteristics from the land-use data. Concerning the purpose inference models, the alternatives consist of specific travel purposes. Attributes that influence the utility of purposes consist of travel characteristics and land-use data near the used stops. The land-use data values are averaged around the used stops, and therefore might not be consistent with the real land-use characteristics at the origin or destination.

Nonetheless, the models estimating the lacking information are handled separately. Combined models might increase the added value, as they consider the complete movement patterns. This ensures consistency of the land-use data with the allocated zones. Due to high complexity and computation times, this approach has not been persevered.

The logit models allow for a disaggregate approach, estimating the origin zone, the destination zone and the travel purpose for individual trips. The specific attributes influencing the allocation are interpretable by the estimated model parameters.

Three different approaches have been applied to construct purpose specific OD matrices based on OV-chipkaart data. The trip-based approach and the tour-based approach use logit allocation models, where the trip-based approach does not take into account the correlation between successive trips, and the tour-based approach does. In addition to these logit models, a straightforward rule-based processing approach has been applied in order to assess the added value of the more complex logit models.

The evaluation of the resulting matrices consists of a series of comparisons between differently constructed OD matrices. Differences consist of the data source, the spatial resolution, the time resolution and the purposes specified. The comparisons are based on linear regression, which indicates overall comparability of OD matrices but does not include structural similarity. Evaluation by means of a MSSIM index has not been pursued since the added value is unknown. A network assignment of the OD matrices can provide additional assessment options, but is time expensive. This topic of route choice calibration, based on OV-chipkaart data, is considered as an interesting follow-up study, but not feasible within this research.

4 The Amsterdam region case study

This chapter presents the case study that has been performed in order to verify the potential of the proposed methodology of constructing purpose-specific OD matrices. In the previous chapter, this methodology has been introduced, based on the available data sources in The Netherlands. The OV-chipkaart is a national smart card system, containing many different regions, with unique public transport systems. This study was concentrated on the Amsterdam region. Here, we go into further detail on the eventual application of the resulting OD matrices and the available data sets.

The literature review already indicated that the preferred methodology depends on the eventual application. Therefore, the chapter starts with the specifications of the intended application in the strategic transport model of the Amsterdam region: the VENOM model (paragraph 4.1). Subsequently, the availability of OV-chipkaart data for this study is presented (paragraph 4.2), followed by a qualitative comparison between the data sets (paragraph 4.3). Finally, the implications of using this specific case study for the generalizability of the method are discussed (paragraph 4.4).

4.1 Eventual application of OD matrices in the VENOM model

The VENOM model is the strategic transport model of the City Region of Amsterdam (SRA), which has been officially in use since 2012. The model forecasts the travel demand for an average working day in the Amsterdam region. The study area covers the larger metropole area of Amsterdam, which includes adjacent areas that have a large influence on the travels in and around Amsterdam. For example, the city of Almere is not part of the Amsterdam region, but is known to inhabit many commuters travelling to Amsterdam. The models zonal grid is adapted from the Dutch postal code system. The spatial resolution is slightly higher than the PC4 level, which means that PC4 zones are generally divided into several VENOM zones, with smaller zones in more urbanized areas.

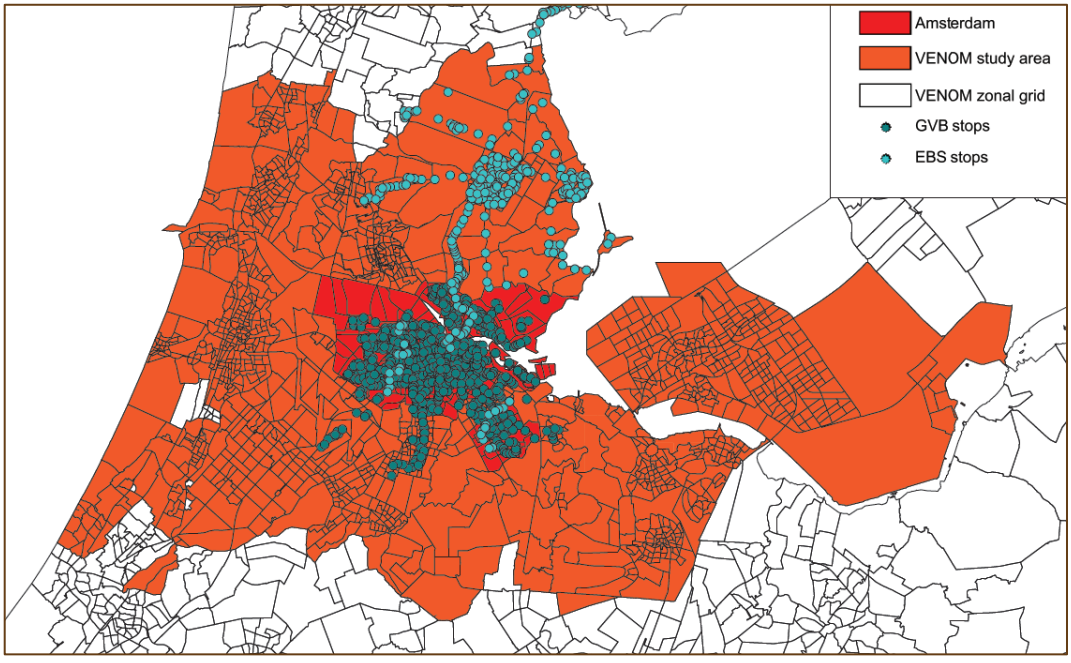


Figure 9: VENOM study area

The VENOM model applies the same pivot point procedure as the main Dutch transport model systems LMS and NRM (Rijkswaterstaat, 2012), but applies a different approach regarding the travel demand of public transport. The VENOM model generates synthetic OD matrices for both train and BTM. These matrices are enriched with survey data because the model does not provide consistent matrices. The matrices for train and BTM are then summed and this public transport matrix is successively assigned to the network. Based on the network loads, the public transport OD matrix is calibrated to improve the fit with cross-section counts (Kieft, Herder, & Pieters, 2013).

The influences of the enrichment and calibration processes on the OD matrix are substantial. During the enrichment the internal trips are increased by approximately 50% for the average working day. During the calibration process, an additional 16% is added to the number of trips. These large alterations invalidate the pursued consistency within the pivot point method. Therefore, the application of OV-chipkaart data for the construction of OD matrices is assumed to provide higher quality OD matrices, although these do not ensure consistency with the synthetic forecast matrices either.

Compared to the OD matrices for road transport, the public transport matrices are not distinguished by purpose. This decreases the model capabilities to specifically implement the influences of policy measures. Furthermore, the currently applied empirical sources for calibration require updating.

This case study initiates several improvements to the currently used OD matrices, which can expand the capabilities of the VENOM model:

- Employment of recent data;
- Employment of high volumes of observed trips;
- Differentiation of matrices by travel purposes.

4.2 Availability of OV-chipkaart data

The responsibility of governance of the Dutch public transport systems lies with regional authorities. The only exception is the concession of the national railways, which is covered by the national government. The SRA is responsible for the procurement of four public transport concessions in its region. This results in the presence of five different public transport concessions in the Amsterdam region[7].

Table 11: Public transport concessions in the Amsterdam region

Concession	*Operator*	*Modes*	*Available*
Amsterdam	GVB	Bus, tram and metro	yes
Waterland	EBS	Bus	yes
Amstelland-Meerlanden	Connexxion	Bus	no
Zaanstreek	Connexxion	Bus	no
National railways	NS	Train	no

[7] It has to be noted that more public transport concessions exist within the study area, since that also includes adjacent regions under control of other authorities than the SRA.

For this study, the data from the concessions Amsterdam, operated by GVB, and Waterland, operated by EBS, have been used. Their area of operation is indicated by their stop locations in Figure 9. This means that the OD matrix of the study area can only be constructed partially. Moreover, concession traversing transfers to the other operators cannot be determined. Since these issues result from the unavailability of data from other operators, rather than limitations of the data itself, we have not pursued enhanced identification of these transfers. When more operators are prepared to contribute to studies that exceed concessionary boundaries, it will be possible to study complete public transport trips. This case study does combine OV-chipkaart data from two different operators, and thereby demonstrates that the proposed methodology of enrichment works for concession-traversing studies. Consequently, this study can stimulate the availability of OV-chipkaart data for future research.

The period for which the travel data is available includes the entire year 2014, for both GVB and EBS data. This long period allows for a longitudinal analysis of the year. However, the vast amount of data also results in long computation times. Therefore, a longitudinal analysis has been performed on several travel characteristics in order to determine a single week, which best represents an average working week. The data of the selected week has been applied in the construction of the OD matrices. The longitudinal analysis is presented in paragraph 5.4.2.

4.3 Matching the WROOV dataset to the OV-chipkaart dataset

The stacked WROOV data of the period 2003 – 2009 contains 279.374 trips within the structural boundaries of the public transport concessions Amsterdam and Waterland. In order to comply the dataset with the target population and to remove missing values, the dataset is filtered with the following selections:

- Only weekdays;
- No student cards[8];
- No missing origins and destinations;
- No missing stops;
- No missing travel purpose.

After this selection procedure, 204.041 trips (73%) remain in the dataset, which comes down to an average of nearly 30.000 observations per year. Trips with missing origin, destination or stops are removed from the dataset. Since Biogeme cannot deal with missing values, these cannot be used in the estimation. The removal of trips with missing travel purpose is applied with the aim of equivalent datasets for the zonal allocation and the purpose inference.

Shifting to the tour-based perspective, the remaining total of 204.041 trips can be classified by tours and non-home-based trips. The used survey for WROOV data collection ensured tours can only consist of two trips: an away trip and a return trip. Only 10% of the trips are classified as non-home-based trip, the majority of 90% is part of a tour (see Figure 10)[9].

[8] In general, student cards are not included in the WROOV studies. However, students were allowed to travel with NVB tickets on a reduced fare, in times their student card was not valid.
[9] The trips within tours and single trips add up to 204.033 trips. The remaining 8 trips that complete the dataset to the 204.041 trips are tours with missing values in one of the trips.

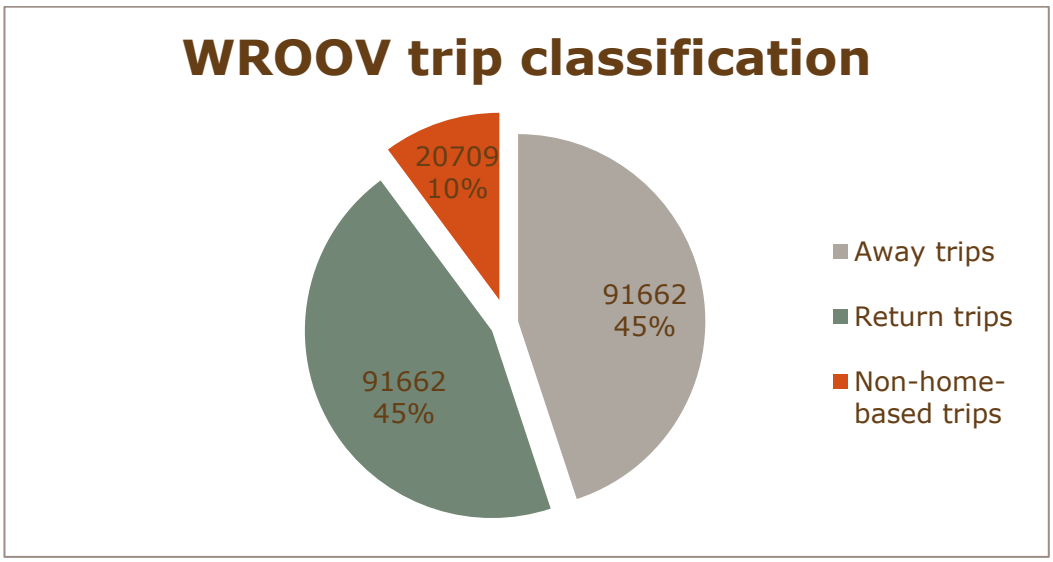


Figure 10: Share of tours and non-home-based trips in the WROOV data

In order to reduce the computation times, for both the trip based as the tour-based modelling approach, a random sample of 40.000 records (20%) is used as training set. This implies that the allocation to zones is independent from time. That is, final model parameter values were expected to be equal for every year within the WROOV data. To validate this assumption, the total dataset is divided per year as well (see paragraph 6.2.4).

Finally, one last selection is applied to the datasets used for the model estimation: the "chosen" zone has to be available in the choice set. From discrete choice theory, we know the choice set should always contain all potential alternatives. Regarding this specific allocation problem, however, it does not make sense to take every possible exception into account. In theory, every zone in the study area is an alternative since it is possible to walk, cycle or drive to any zone. For the construction of base matrices we do not want to consider exceptions where travellers cover large distances to and from the PT stops they use. Therefore, chosen zones that are not within the catchment area of a stop are filtered from the data. Paragraph 6.2.1 provides the details of the choice set generation by means of catchment areas and presents the number of observations used of the model estimation per modelling approach and choice set generation method.

While the Biogeme software does provide the option to apply weights on observations, the weights available in the WROOV data are not applied in the estimation. These weights are based on the division of the revenues of ticket sales and are not constructed with the goal to represent the origins and destinations of travellers. In addition, the selection procedures that have been applied on the original WROOV dataset invalidate the weight factors (see Appendix A). It is possible to compensate for the applied selections, but because of the complex and unclear determination of these factors, we decided to use the observations without weights.

4.4 Generalizability of the case study

With this case study we aim to demonstrate the value of the method of enriching smart card data. In order to determine the implications of using this specific case study, this paragraph discusses the generalizability of the case study, concerning the

structural boundaries due to limited availability of data and the eventual application of OD matrices.

This study focussed on the travel with bus and light rail, where a complete description of the travel with all public transport modes would have been more valuable. However, for both the WROOV survey and the OV-chipkaart, no data concerning travel by train was available. Regarding OV-chipkaart data, this is a result of unavailability of data, in contrast to WROOV, which does not cover train travel.

Amsterdam is a unique situation for public transport in the Netherlands, which may comprehend significantly different travel patterns compared to other regions. Only the larger cities in The Netherlands provide public transport by light rail. Furthermore, Amsterdam attracts high volumes of tourists, both domestic and foreign. These factors might result in different model parameters. The estimated enrichment models for the case study, therefore, may not be applicable for other regions. Nonetheless, the WROOV data provide the opportunity to estimate the models for specific regions, since it covers the entire county. Moreover, the OV-chipkaart data structure is equal over the entire country. Hence, a similar procedure is applicable for other regions.

Besides different travel patterns and utilization of the public transport system, also differences in the smart card system require consideration when relating this case study to application of smart card data in travel demand studies abroad. The OV-chipkaart registers both boarding and alighting stop, as well as transaction locations. For ticketing system based on flat fares or systems without an integrated GPS system, techniques are available to infer the required information to employ the methodology of this study (see paragraph 2.2). However, the data quality will be less for inferred attributes than observed attributes.

The aimed application of OD matrices in the VENOM model is limited due to the data constraints. The fact that VENOM applies combined public transport matrices increases the limitations due to the unavailability of train data. On the other hand, the resulting matrices, based on data from the concessions Amsterdam and Waterland might also be valuable for the Amsterdam City Model (VMA).

5 Public transport travel analysis

This chapter presents the results from the travel analysis that has been performed on the WROOV and OV-chipkaart datasets. The literature review has provided an indication of attributes relevant to the information we want to add to the OV-chipkaart data. Chapter 3 described the data sources and which attributes are applicable as key variables in the model estimation. These attributes have been examined in the WROOV data in order to provide an indication of their predictive value. In addition, characteristics which are not available in OV-chipkaart data, but could have a possible predictive value, have been examined to comprehend the overall predictive value of the key variables. Furthermore, the stability of the information to be added has been investigated in order to derive the durability of their predictive value. Subsequently, a comparison of the datasets on key variables provides insight in the suitability of key variables as predictors. These classifications of predictors form the foundation of the model estimation, described in the next chapter.

First, the lacking information in OV-chipkaart data is analysed in the WROOV data:

- the access and egress trip legs are expounded as a foundation for the conversion of stop-based matrices to zonal matrices (paragraph 5.1);
- a depiction of the distribution of travel purposes provides input for the estimation of the purpose inference models (paragraph 5.2);
- the description of concession traversing transfers provides information for the filtering of trips with origins or destinations outside the study area (paragraph 5.3).

Then, after these analyses of WROOV data, a comparison with OV-chipkaart data on key variables is presented (paragraph 5.4). The chapter is concluded by a summary of the findings from the data analysis (paragraph 5.5).

5.1 Access and egress trip legs

In paragraph 1.4 we introduced different definitions to describe the trip ends: the trip-based *origin and destination* or the tour-based *home-end and activity-end*. In the analysis of access and egress trip legs, the tour-based definition provides more insight through the additional information it contains. Since the WROOV data set contains mostly tours with two trips (see Figure 10), the access and egress trip legs are registered once at both ends, resulting in almost equal statistics for both access and egress trip legs in the trip-based definition. Considering the tour-based definition, access and egress trip leg statistics are specific for the home-end and activity-end. Hence, the trip-based definition allows for analysis on access and egress distances combined as one phenomenon, where the tour-based approach allows for analysis of the specific trip ends.

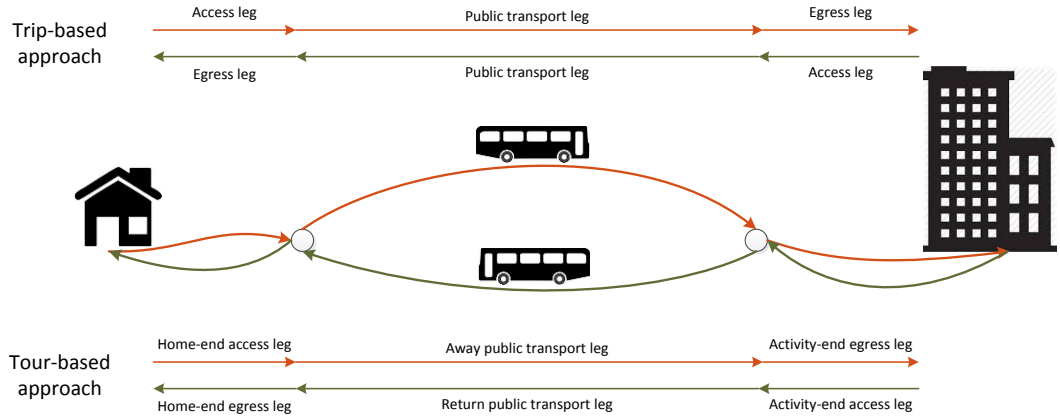


Figure 11: Trip-based and Tour-based definitions of trip legs

Origins and destinations, as well as the boarding and alighting stops are registered in the WROOV dataset. Origins and destinations are available at the level of PC6 zones, which are not equivalent to addresses, but comparable to housing blocks. These small zones are represented by the coordinates of their centre. Stops are represented by the coordinates of the stop cluster, which means that the clustered stops in opposite directions of the same line are represented by the same coordinates. Access and egress distances are calculated by means of the Euclidean distance between these coordinates, which is not equal to actual walking or cycling distance, as these depend on the local infrastructure. Hence, the access and egress distances considered in this study are a slight underestimation[10] of the actual distances travelled.

Data cleaning has been applied to reduce the effects of wrongly registered origins and destinations, resulting in approximately 13% of the data to be excluded for the analysis of access and egress trip legs. The data cleaning process is described in Appendix A. However, some erroneous data remains, with very large access and egress distances. Most likely, these are caused by switched origins and destinations. Since approximately 93% of the distances are below 1500 metres, we focus on this interval in the analysis in the remainder of this paragraph.

5.1.1 Key variables

First, we have aimed to relate the access and egress distances to attributes available in OV-chipkaart data. Previous studies (Utsunomiya, Attanucci, & Wilson, 2006) (Alshalalfah & Shalaby, 2007) have indicated that the access and egress distances depend on the level of service provided at the considered stop. The level of service is described by the frequency, the speed and the directness of the transport service. Travellers are prepared to walk of cycle further for transport with higher speeds and higher frequencies. Since these characteristics are not readily available, we investigated the distances by mode, as these pertain different levels of service.

The analysis shows different distributions of access and egress distances between modes. On average, travellers cover longer distances to and from metro stops. The difference on the home side is larger than on the activity side. Moreover, on the home side travellers cover larger distances to and from tram stops compared to bus stops, while this deviation is not observed at the activity end.

[10] Assuming a rectangular infrastructure pattern, the maximum underestimation is a factor $\sqrt{2} \approx 1.4$

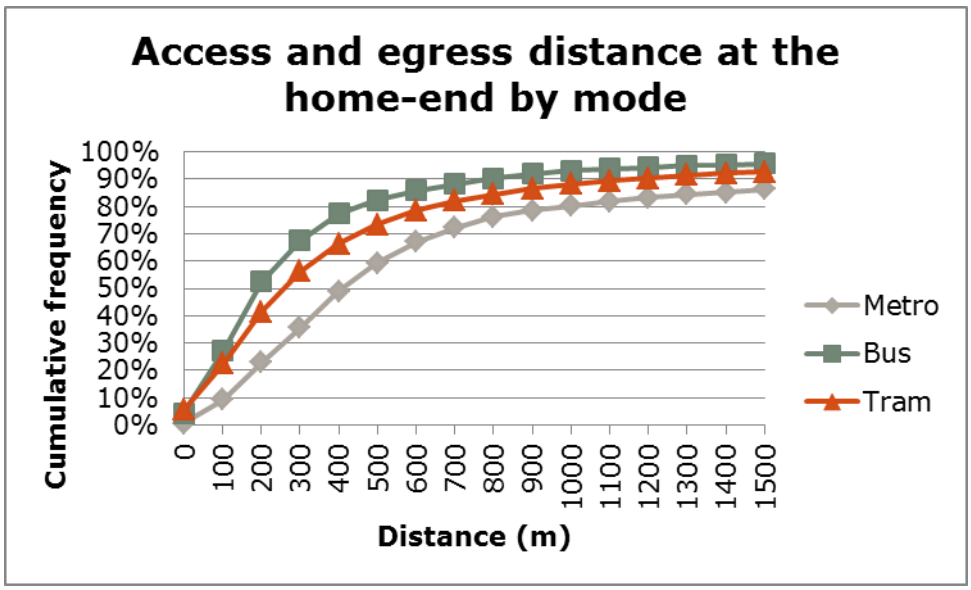


Figure 12: Access and egress distance distributions by mode at the home-end

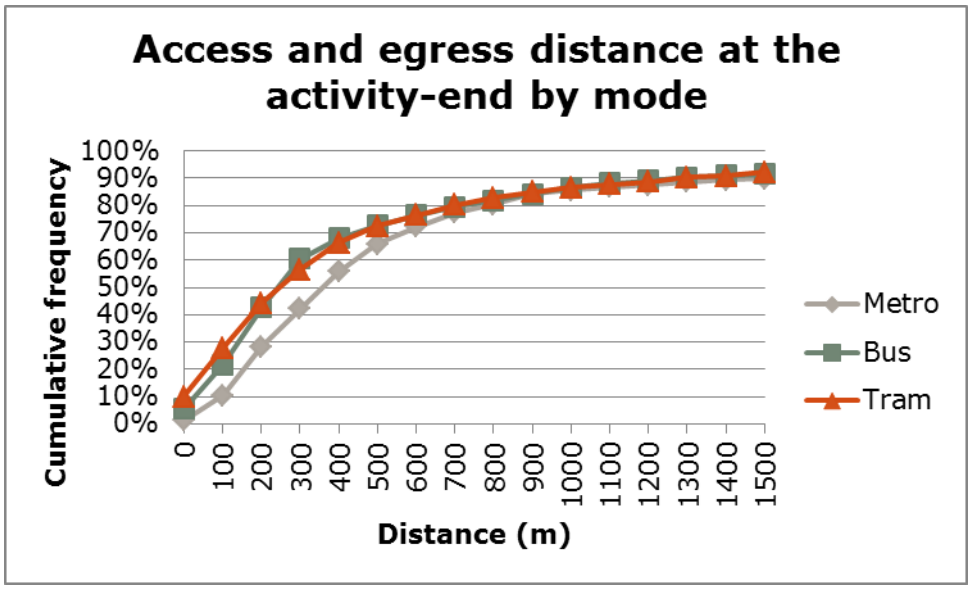


Figure 13: Access and egress distance distributions by mode at the activity-end

Besides the mode, we also expected a relation between the degree of urbanization and the level of service. In more urban areas, travellers have more options in their stop and route choice, while in more rural areas, travellers are more likely to have lesser options and can be classified as "captives" of the network, and consequently need to cover longer access and distances. However, the analysis of the degree of urbanization at the used stop does not present a clear correlation. While on average longer distances are covered in very rural areas (class 5), in very urban areas (class 1), distances are similar to other degrees of urbanization. This might be caused by the fact that the metro system only serves highly urban areas. Again, differences are only observed at the home-end and are less apparent at the activity-end.

For travel distances below 15 km, no differences are observed between the home-end and the activity-end. However, for longer travel distances, the average access and egress distances decrease, where they remain equal at the activity-end.

Distributions of access and egress variables with additional key variables are presented in Appendix B.

5.1.2 *Characteristics unavailable in OV-chipkaart data*

The influence on the access and egress distances of three personal traveller attributes, which are not available in OV-chipkaart data, has been investigated: the travel purpose, the gender and the age. The results of these analyses provide additional insight in the relative explanatory value of the key variables. Graphs related to these analyses can be found in Appendix B.

From these analyses we conclude that the gender is not related to the access and egress distances on either end of the trip. Regarding the travel purpose and the age of the traveller, differences in access and egress distances are observed at the activity end. Children and especially senior travellers cover shorter distances on average between the used stop and their activity location. Furthermore, trips made for the purpose of *shopping*, and to lesser extent the purpose *other*, contain shorter access and egress legs. The can be explained by two possible phenomena. Either travellers are prepared to cover larger distances for the purposes *work* and *education*, or shopping locations are better served by public transport than offices and schools.

At the home end, access and egress distances do not differ for different travel purposes or ages, which indicates that the access and egress distances are mainly influenced by the level of service.

5.1.3 *Longitudinal analysis*

The access and egress distances are stable over the WROOV years. Both the trip-based definitions and the tour-based definitions only have slight variation over the years. Figure 14 shows the 75 percentile of the access and egress distances, since the mean value is to a larger extent influenced by erroneous data.

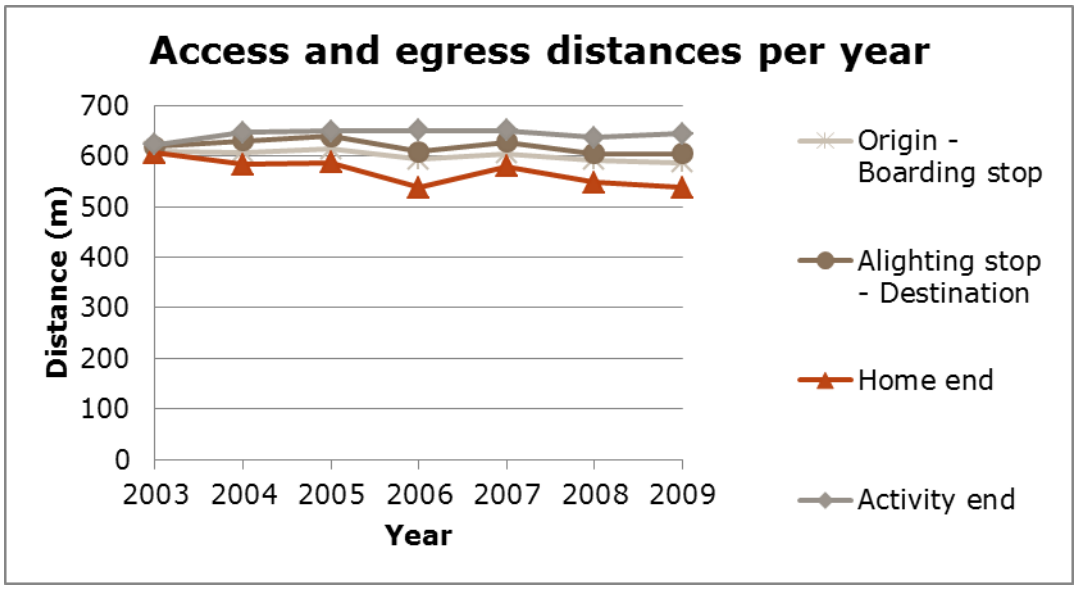


Figure 14: Longitudinal analysis of access and egress distances

5.2 Travel purpose

For the identification of relevant attributes for the purpose inference, a similar method has been pursued as with the access and egress distances. First, the relation with key-

variables has been investigated. Second, to place the explanatory value into context, unavailable attributes in OV-chipkaart data have been analysed. It has to be noted that the relative frequencies per purpose are based on shares per purpose, hence adding to 100% for each purpose. This does not relate to the absolute frequency of each purpose. Figure 15 presents the overall shares of the four identified purposes.

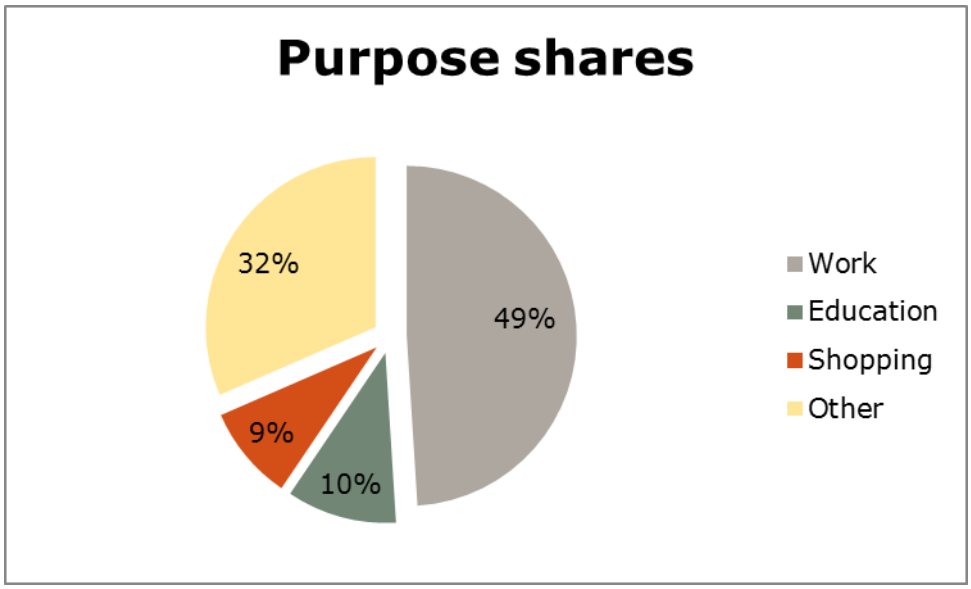


Figure 15: Overall purpose shares in the WROOV data

5.2.1 Key variables

The literature indicates a clear relation between the activity duration and the travel purpose. Therefore, it was applied as the main attribute in the rule-based processing approach for purpose inference by (Chakirov & Erath, 2012).

The distribution of activity durations per purpose in WROOV data confirms the indicated relevance of the travel purpose. Clear peaks are visible for the purposes work, between nine and ten hours, and education, between seven and eight hours[11]. The purposes shopping and other show less sharp peaks in activity duration, but consist of mostly activities shorter than six hours.

[11] Note that the activity duration includes the travel time of the away trip due to the lacking alighting time in WROOV data.

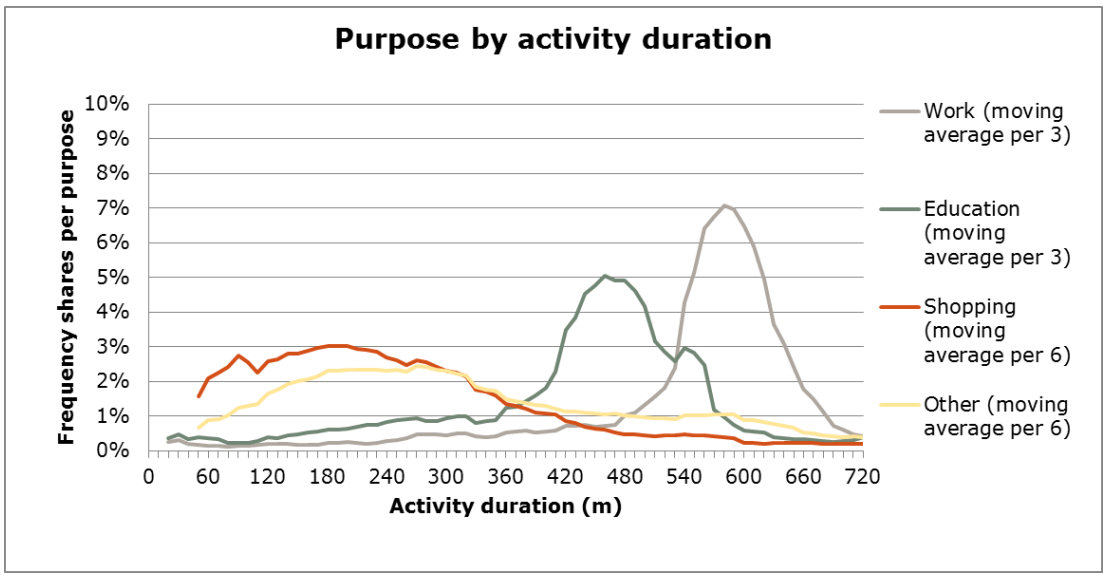


Figure 16: Activity duration distributions per purpose[12]

The second-most attribute related to the travel purpose in the literature is the departure time. The shares of departure time per purpose show similar patterns for the purposes work and education. Both purposes have strong peaks in the morning and afternoon, with the distinction that the afternoon peak of education trips is earlier than the afternoon peak of work trips. Furthermore, the departure time patterns of the purposes shopping and other are similar. These purposes mostly occur during the day, between the morning and afternoon peaks. A distinction is that the purpose other is relatively frequent in the night.

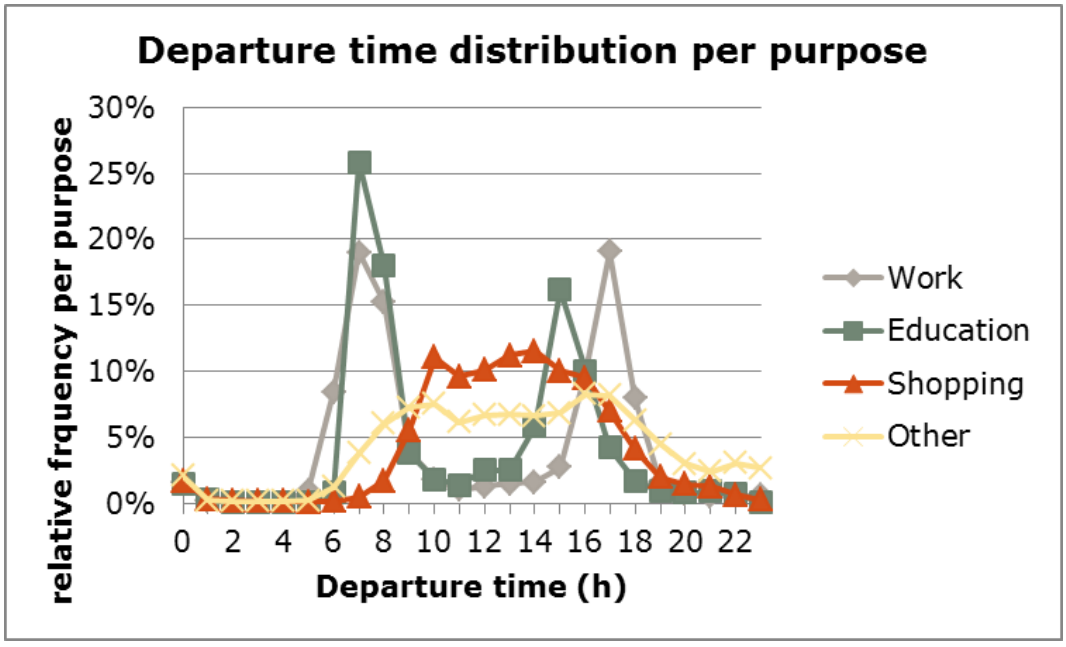


Figure 17: Departure time distributions per purpose

A third attribute indicated as relevant for the travel purpose in literature is the travel frequency. The travel frequency distribution shows a similar distinction between the

[12] Moving averages are applied in order to correct for rounding of times by respondents. The observed frequencies are obtainable in Appendix B.

compulsory travel purposes *work* and *education* on the one hand, and the discretionary purposes *shopping* and *other* on the other hand.

In addition to these three attributes frequently identified as possible explanatory variables for the travel purpose, the distributions of further key variables per purpose have been investigated. Many of these attributes show different distributions per purpose, and therefore are potential attributes for the purpose inference. Here, only the distribution over contracts is presented, since it shows the largest distinction between purposes. In Appendix B distributions are presented for the travel frequency, travel distance, number of legs per trip, operators used, and the product fares.

The distribution of contract types per purpose does show a distinction between the purposes work and education. Travellers with the purpose work mostly used year contracts, while travellers with the purpose education mostly used monthly contracts. On the other hand, the purposes shopping and other mostly travel without a contract. The distribution of travel purposes per fare show a very similar distribution as the contract durations, indicating a high correlation between the attributes.

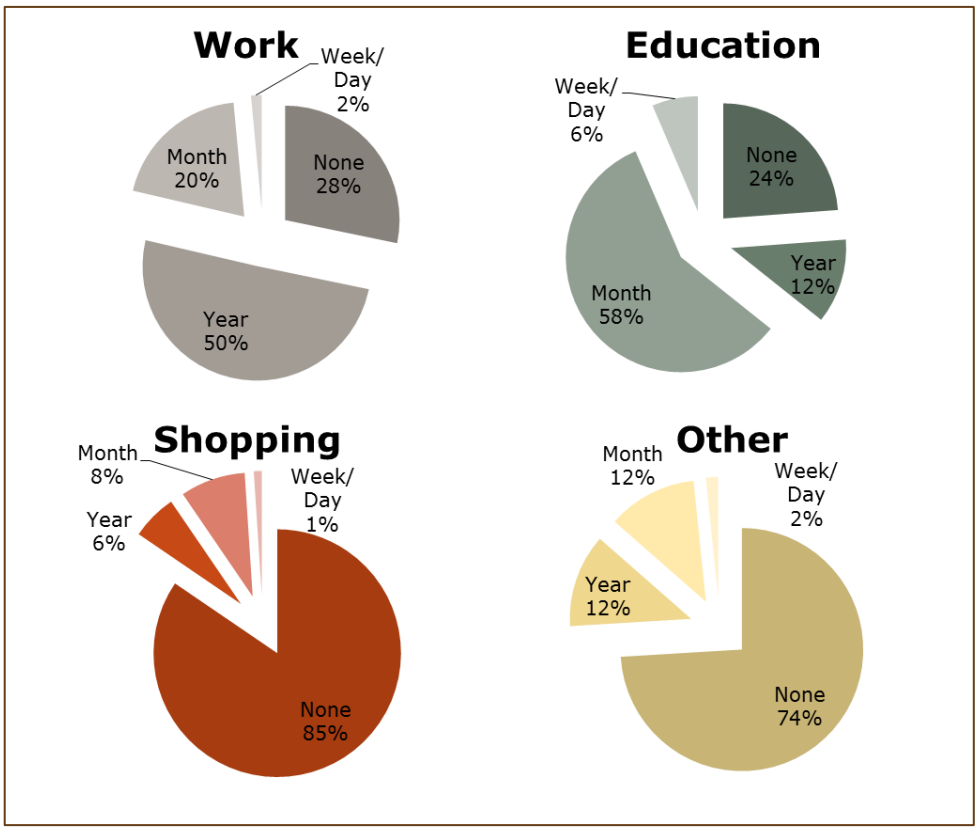


Figure 18: Distribution of contract types per purpose

The distributions of travel distances per purpose only show slight differences between the purposes. However, looking at the relative purpose shares over the travel distance, the share of the purpose work steadily increases with longer distances. The purpose shopping has a strong peak for short distances, but also longer trips are made for shopping. This indicates the difference of shopping for daily groceries and shopping as a recreational activity, which are both included in this category.

5.2.2 Characteristics unavailable in OV-chipkaart data

The OV-chipkaart data does not include information on the gender of the travellers, but in case of personal cards, it does contain the age of the traveller. This information has not been used in this study to reduce the possibility of privacy violations. Moreover, the age sample of personal cards is not considered as representative for the entire system. Nonetheless, the correlation between age and purpose has been investigated, since it is considered as a valuable predictor.

The distribution of age groups per purpose confirms the expectations. The purpose *work* is mostly applicable to adults, while *education* trips consist of mostly children. Seniors have low shares of trips in these compulsory purposes and mostly travel for *shopping* or *other* purposes. Hence, it can be concluded that the age would be a valuable estimator for the purpose inference.

5.2.3 Longitudinal analysis

The longitudinal analysis of the purpose distribution shows a slight increase for the share of the purpose work over time. It has to be noted that the WROOV data of 2009 do not cover the entire system due to the partial implementation of the OV-chipkaart. The e-purse travel was implemented in that year, replacing the *strippenkaart* tickets, while contracts were still paper tickets, included in the WROOV study. Furthermore, the year 2003 shows a slight distinction with the trend, which can be explained by initialization of the study.

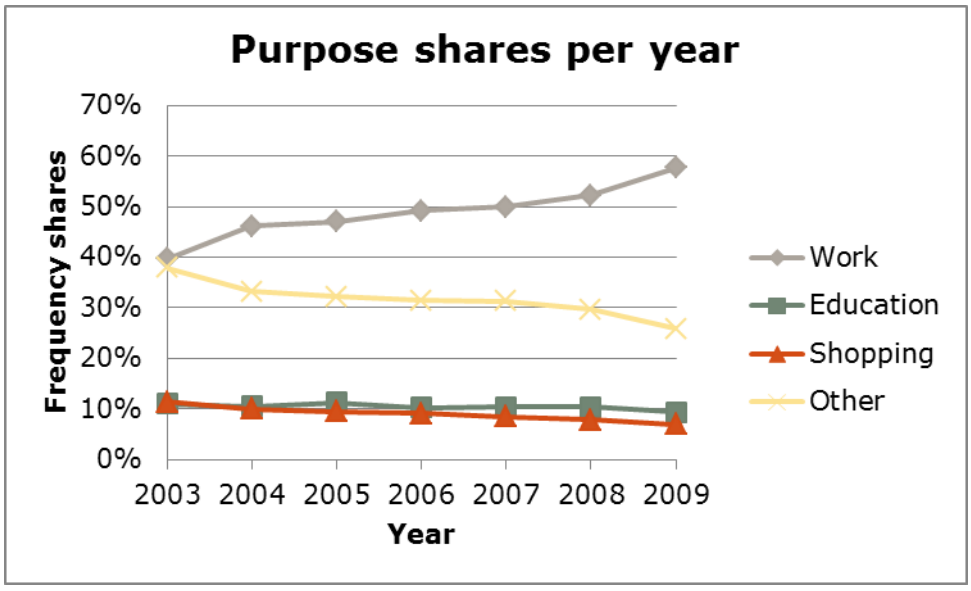


Figure 19: Longitudinal analysis of the purpose shares in the WROOV data

5.3 Concession traversing transfers

Since we do not have the data from all public transport operators in the study area available for this study, it is not possible to derive transfers to all operators from the data. These transfers do have an impact on the OD matrix construction, as the origins and destinations are outside the catchment area of the stop where the transfer is made. Since this problem is not part of the research focus, but a consequence of the unavailability of data for this study, we have aimed at a straightforward solution. When data from all operators would be available, this problem would cease to exist.

The WROOV data contain data for all BTM operators, but not for the national railways. However, respondents did indicate whether a trip by train was made before or after their travels with bus or light rail. Although this indication is not irrefutable, as it was not a specific question, it provides an overview of the share of travellers transferring to and from the train. These shares have been analysed per train station and are presented in Table 12.

Table 12: Shares of BTM trips transferring to or from the train network per train station

Train station	*Trips*		*Tours*		*Non-home-based trips*	
	Boarding	Alighting	Home-end	Activity-end	Boarding	Alighting
Amstel	31%	31%	50%	12%	32%	16%
Bijlmer	13%	14%	33%	8%	14%	7%
Centraal	40%	39%	61%	15%	41%	17%
Diemen	23%	34%	10%	60%	0%	20%
Diemen Zuid	14%	15%	17%	11%	14%	9%
Duivendrecht	75%	74%	85%	39%	69%	30%
Holendrecht	9%	10%	8%	9%	11%	12%
Lelylaan	23%	23%	26%	19%	26%	11%
Muiderpoort	16%	17%	20%	10%	14%	11%
Purmerend	34%	33%	26%	36%	25%	20%
Purmerend Overwhere	11%	11%	5%	33%	4%	6%
RAI	19%	19%	31%	10%	29%	12%
Schiphol	33%	32%	65%	17%	24%	13%
Sloterdijk	53%	52%	76%	17%	51%	21%
Weesp	45%	46%	51%	38%	41%	15%
Wormerveer	40%	37%	35%	42%	20%	21%
Zuid	32%	31%	59%	12%	40%	17%

The shares of transfers at train stations relate to the shares of trips that use BTM as access transport, in case of a transfer at the alighting stop, or egress transport, in case of a transfer at the boarding stop. The fact that the metropole area of Amsterdam has fourteen train stations indicates that the train system also has a regional function. The transfer shares also prove that the train system and the BTM systems are closely related. The shares of trips used for access or egress transport are very high for the stations Duivendrecht, Sloterdijk and Centraal. These stations are very well connected to the GVB network, which explains these large shares.

The transfers shares deviate substantially between stations and, moreover, between the home-end and the activity-end. In general, train stations in Amsterdam indicate higher transfer shares at the home-end compared to the activity-end. Train stations

outside Amsterdam show a different picture, with in some cases even an opposite relation.

5.4 Quantitative comparison of key variables

After determining the relation of specific attributes to the travel purpose, their appropriateness as medium of the information transfer has been assessed by means of a quantitative comparison of the key variables of both sources. In order to create a comparable dataset from OV-chipkaart data, however, rule-based processing had to be applied for the identification of trips. Therefore, we discuss these processing rules first and the applied OV-chipkaart dataset first, before we present the results of the quantitative comparison.

5.4.1 Construction of trips and tours from OV-chipkaart data

In order to compare the WROOV data with OV-chipkaart data, the raw OV-chipkaart data had to be interpreted. This included the application of processing rules to distinct transfers and activities between consecutive trip legs, which influenced the key variables. Figure 20 contains the flow diagram with the applied processing rules. The complete overview of the data handling procedures is included in Appendix A.

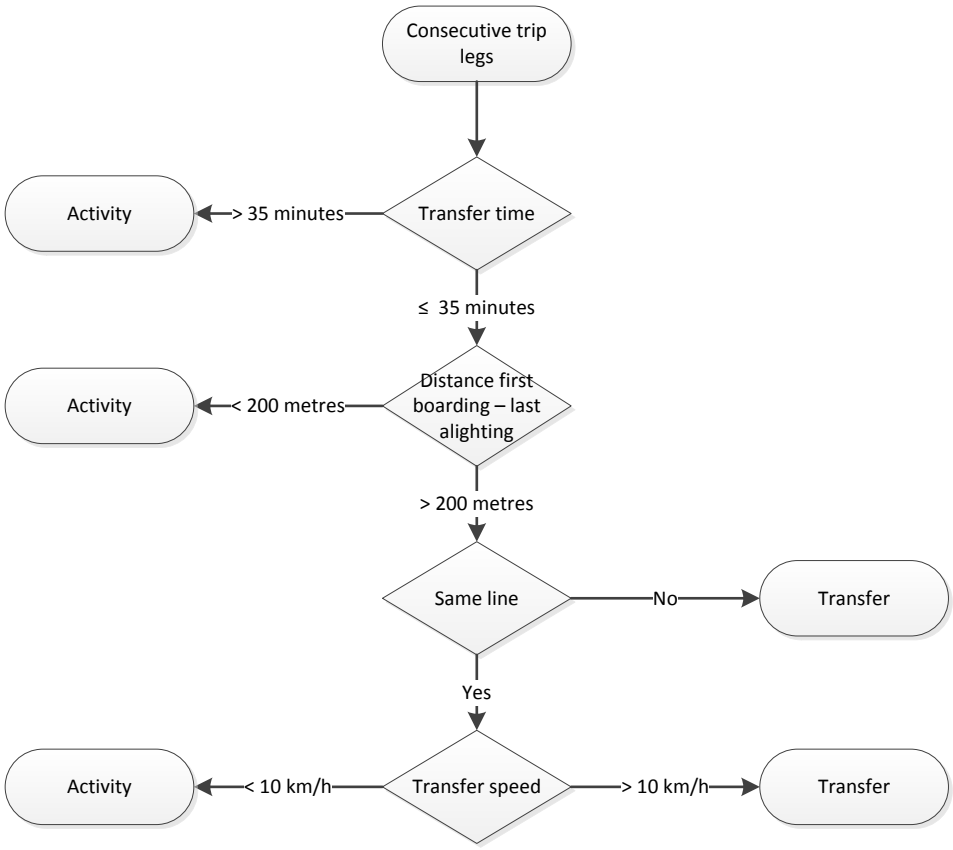


Figure 20: Flow diagram of distinction between transfers and activities

The distinction between transfers and activities is based on constraints in the three dimensions indicated in the literature (see paragraph 2.2.2): time, space and structure. The transfer time is the primary indicator and includes the constraint of 35 minutes commonly applied in Dutch public transport (Nijënstein & Bussink, 2014). This time constraint is brought about by the assumption that transfers options are always available within this time window, hence it is also related to the public

transport structure. The second indicator is related to the identification of tours. When two consecutive trips consist of an away trip and a return trip, thus when a tour is made, an activity is inferred, regardless to the transfer time. Therefore, the spatial constraint of 200 metres between the first boarding stop and the last alighting stop has been applied as indicator of a return trip. The third indicator is a structural constraint: an activity is inferred if a transfer is made on the same line. Devillaine and colleagues (2012) argue that a transfer on the same line, in both directions, implies an activity. However, the analysis of transfers on the same line in OV-chipkaart indicated that traveller's check-in and check-out in the same vehicle, while driving. In order to correct for this fare-dodging behaviour, we implemented an additional constraint of the transfer speed for transfers on the same line. The basic application of only the time-constraint of 35 minutes results in an overestimation of transfers by 22% compared to the enhanced distinction between transfers and activities.

5.4.2 Longitudinal analysis of OV-chipkaart data

The second step toward a quantitative comparison between WROOV and OV-chipkaart data, is the selection of a data collection period. In strategic transport modelling, common practice is to use an average working day as capacity of the travel demand. This average working day does not actually exist, as it is an average, but is does provide a more stable measurement, which contemplates with the requirements of long-term forecasts. The availability of a full year disaggregated travel data has provided the unique opportunity to place an average working day in context. Moreover, this allows for a more deliberate "construction" of the average working day.

The total number of trips per week clearly depicts the holiday periods and, less obviously, the seasonal variance. Excluding the weeks with holidays, the average travel demand of work weeks looks constant, with slightly decreased demand in the spring and slightly increased demand in winter. From this, we conclude that the construction of an average working day does not need to be based on a long period but can be based on a single week, which drastically decreases the computation times.

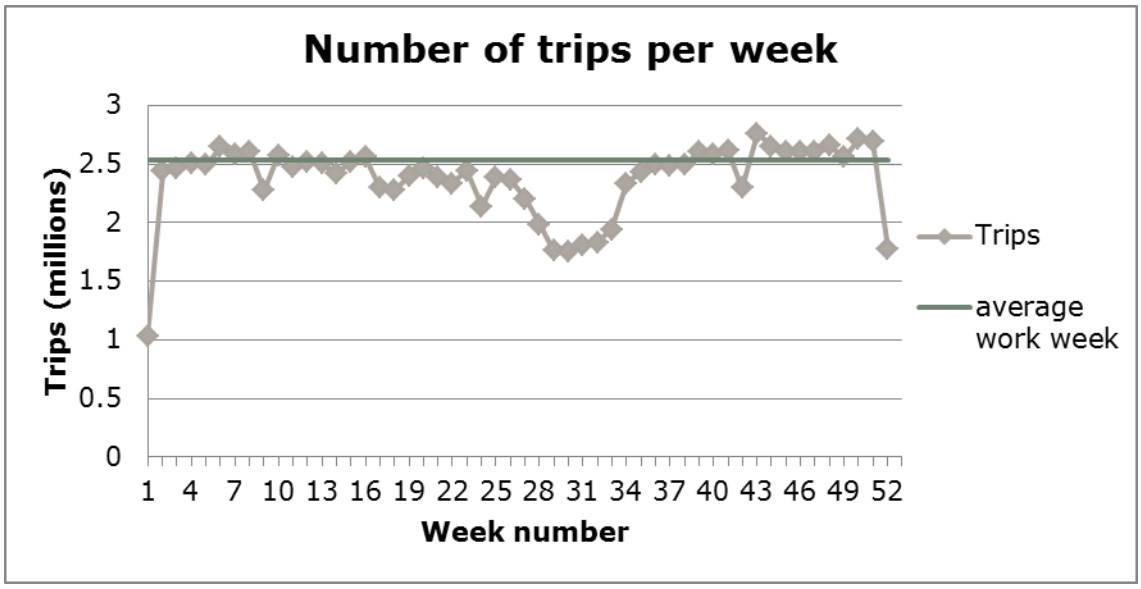


Figure 21: Trips per week in the OV-chipkaart data

Analysing the stability of key variables in OV-chipkaart data, we have found that these are rather stable as well, again with the exception of holidays. This indicates that

these variables are appropriate for enriching the OV-chipkaart data in context of an average working day.

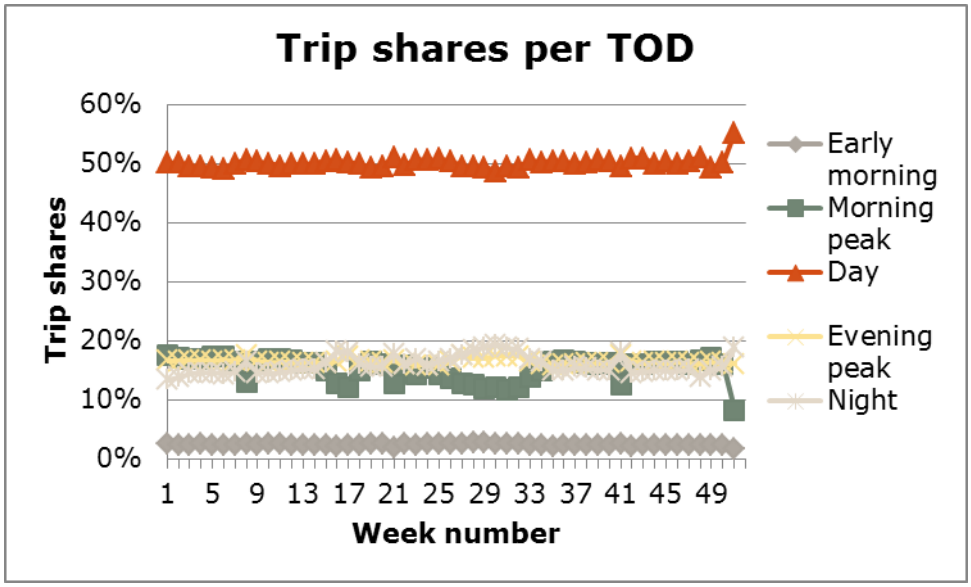


Figure 22: Trip shares per time of day over the year

The overall statistics are fairly constant over the year, with exceptions of holiday periods and national days. For computational reasons, a selection of 1 week has been made to construct the average work day.

Week 49 is closest to the average work week; hence this week has been used in further analyses and for the construction of the OD matrix for an average working day. This one week of crude OV-chipkaart data has been processed, resulting in a trip dataset, representing the average working week of 2014. An overview of the data processing procedures can be found in Appendix A.

5.4.3 Characteristics related to access and egress trip legs

Based on the selected week, we have compared the description of the travel demand by the OV-chipkaart with the description of travel demand by WROOV. Ideally, the distribution of key variables agrees between the sources. In that case, the key variables are appropriate for transferring the information between the sources. In order to match the target populations of the OV-chipkaart dataset and the WROOV dataset, students (20%) and short term contracts (9%) are filtered from the data, since these are not included in the WROOV survey.

The key variable most related to the access and egress distance is the mode. The comparison of both sources on the trip distribution over modes shows an overestimation of the trips made with multiple modes by the WROOV data. The modal shares of bus are very similar, so the overestimation of multiple modes is at the expanse of tram and metro trips.

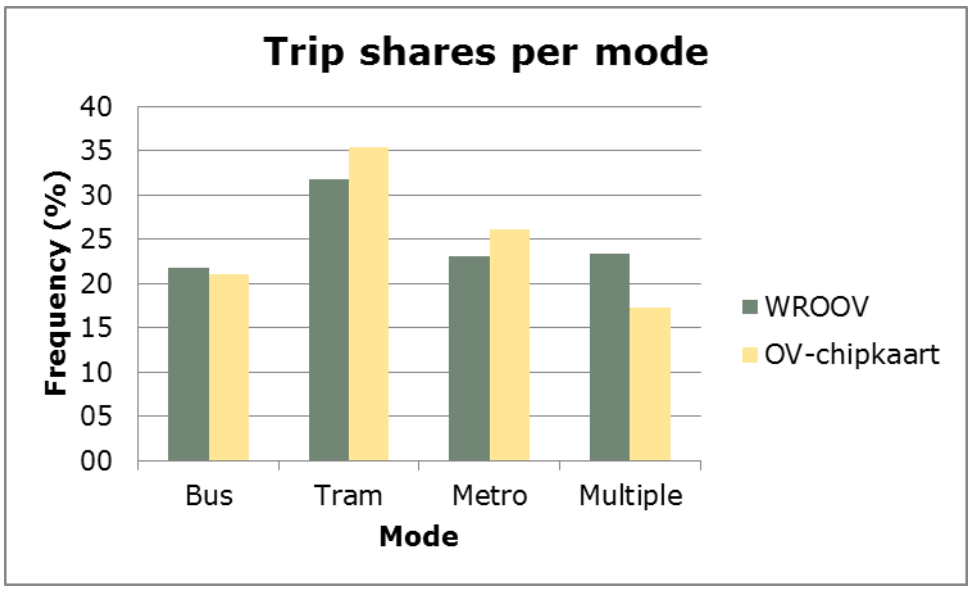


Figure 23: Comparison of trip shares per mode

5.4.4 Characteristics related to the travel purpose

The mode is also correlated to the travel purpose, showing a higher share of work trips for the modes metro and multiple modes. Consequently, WROOV over represents the purpose work.

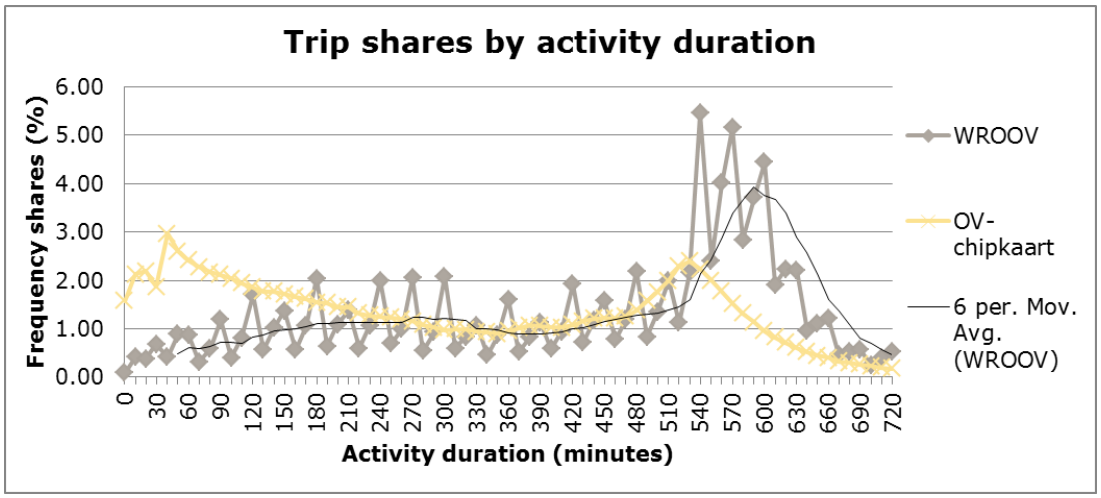


Figure 24: Comparison of trip shares over the activity duration

The comparison of activity durations in the sources indicates an overrepresentation of long activities in the WROOV data. Consequently, work trips are overrepresented. In addition, short activities are underrepresented in the WROOV data. This is a well-known problem of travel surveys due to respondents that forget to report short activities or think these are not important. The distribution of activity durations in OV-chipkaart data also shows a clear peak at the 35 minute interval. This indicates that the applied processing rules for the distinction between transfers and activities do not catch all activities. As a result, short activities are underrepresented by OV-chipkaart data as well. On the other hand, the applied processing rules also indicate activities close to zero minutes. This is not realistic and probably caused by errors in the processing rules and errors in the time recording.

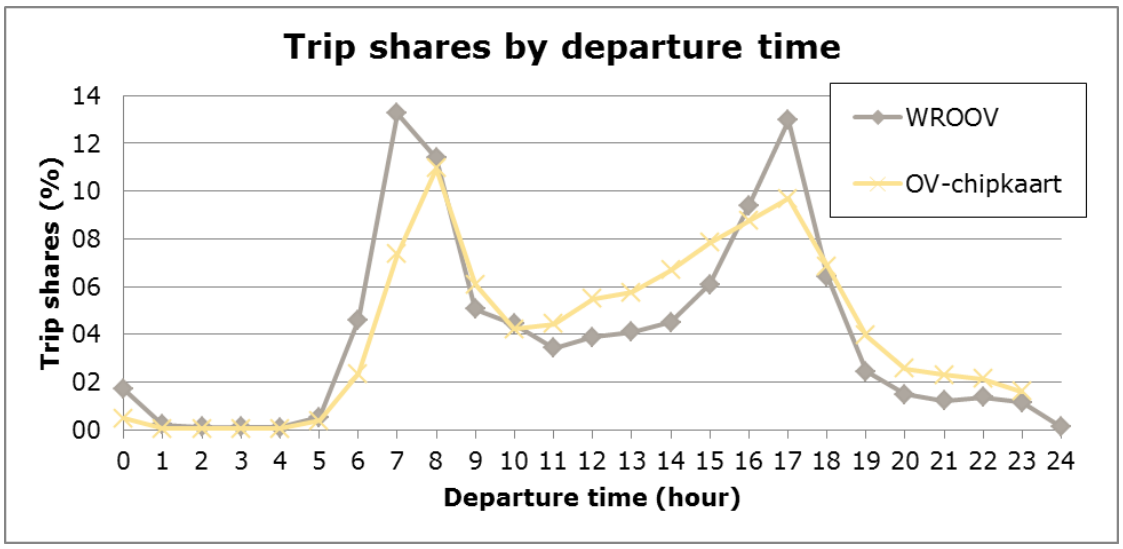


Figure 25: Comparison of trip shares over departure time

The distribution of trips over the day shows clear peaks in the morning and evening in both sources. However, the peaks are sharper in the WROOV data compared to the OV-chipkaart data. The OV-chipkaart shows a larger share of trips during the day, between the peaks, and also during the night. This indicates a larger share of discretionary purposes.

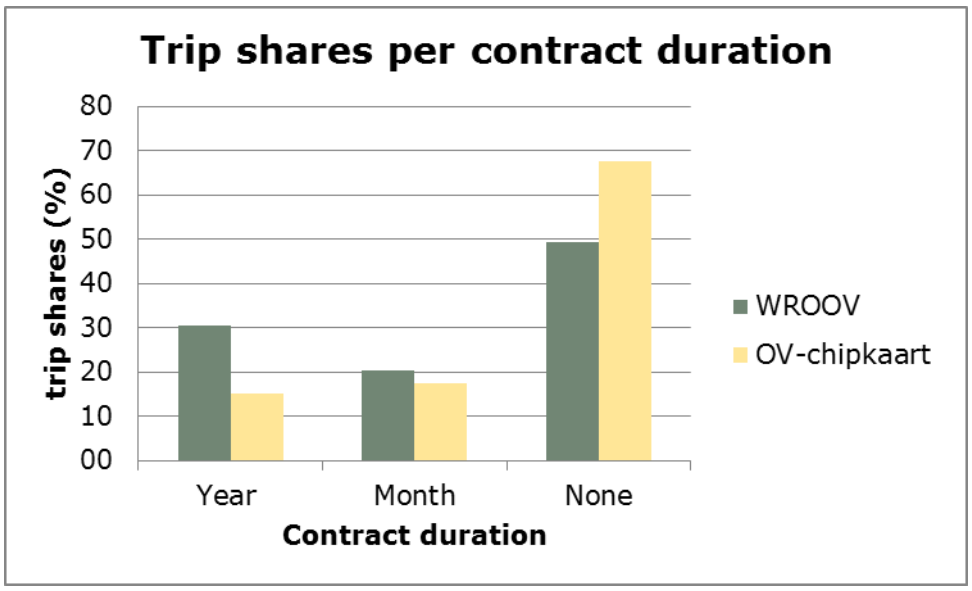


Figure 26: Comparison of trip shares per contract duration

The distribution of contract durations also dissimilarities between the sources, as the WROOV data over represent trips made with contracts, mostly year contracts. The OV-chipkaart has a substantially larger share of trips made without a contract. As a result, WROOV over represents work trips and, to lesser extent, educational trips.

The WROOV surveys also over represent trips with higher frequencies, showing many tours that are made four or five times a week, where the OV-chipkaart shows mostly tours made only once or twice a week.

The share of trips per concession in WROOV also shows a deviation with the OV-chipkaart data. WROOV slightly over represents the share of trips made in Waterland.

Overall, we can conclude that most of these imbalances are related to the same problem: the larger share of contracts in the WROOV data. Consequently, WROOV represents a larger share of *work* trips. The higher share of trips outside the peak periods, without contract and with shorter activities can also be caused by the insufficient filtering of international travellers from the OV-chipkaart dataset, as these can also travel with a regular OV-chipkaart with stored value.

5.5 Conclusions regarding the travel analysis

The travel analysis had two main goals: the identification of relevant attributes for the purpose inference model and the identification of attributes that can be applied in the differentiation of catchment areas for the zonal allocation models.

5.5.1 Predictive value of attributes

The literature relates the level of service to the access and egress distances to public transport stops. However, line frequencies and travel speeds are not available in the WROOV data. Therefore, the mode of transport has been indicated as the main key variable in order to distinguish between catchment areas of stops, since it attains different levels of service for each mode.

In addition, differences were found between the access and egress distances at the home-end and the activity-end. At the home-end, access and egress distances seem to be mainly related to the level of service. Conversely, on the activity end, the access and egress distances show higher correlation with the travel purpose. These differences are related to the geographical location of the used stops and therefore are not considered as a valid distinction between access and egress distances at specific public transport stops.

The variance in the access and egress distances cannot be fully explained by the available attributes in the data. Even between stops with the same modes, large differences are observed in the distribution of access and egress distances. In order to obtain more insight in the catchment areas of stops, it is advisable to include the number of lines serving a stop, their frequencies and their operational speeds.

Considering the attributes related to the travel purpose, several key variables are indicated as potential explanatory variables for the purpose inference. Besides the frequently mentioned attributes of activity duration, departure time and frequency, also the contract duration, the fare and the distances travelled are correlated to the travel purpose. The distribution over key variables shows a strong correlation between the purposes work and education, with similar patterns over the departure time and frequency. In addition, the purposes shopping and other show a correlation in the distribution over these attributes as well. Attributes that do indicate differences between compulsory and discretionary purposes are the activity duration and the contract duration. Another relevant, but unavailable attribute is the age of the traveller.

5.5.2 Appropriateness of key variables as medium of information transfer

All differences that have come to light in the quantitative comparison of both sources can be associated with the higher share of contracts in the WROOV data compared to OV-chipkaart data. This dissimilarity can probably be explained by the coverage of

WROOV survey, which did not include regional tickets or international respondents. Since the popularity of Amsterdam as tourist attraction has only increased in recent years, together with the indirect filtering of tourists based on short term contracts, this leads to difference between the samples of the two sources.

Travellers with contracts tend to travel more for compulsory purposes *work* and *education*, whereas travellers without contract mainly travel for the discretionary purposes *shopping* and *other*. Hence, estimation of the influence of key variables based on WROOV data and application of these estimates on OV-chipkaart data might result in biased estimates due to differences between the data sets. This problem can only be solved by estimating the influence of model attributes based on a representative sample of the current travellers. This is not a feasible solution within this research due to budget constraints, but might be required in future years for the durability of this method.

6 Estimation of enrichment models

This chapter describes the process and presents the results of the model estimation for the three allocation subjects. First, the construction of reference matrices is described (paragraph 6.1), constructed by rule-based processing. Then, the zonal allocation models are described (paragraph 6.2), including the model structure, choice set generation, available attributes, model enhancement and the final parameter estimates. The allocation models for both trip-ends are closely related and therefore presented collectively. Subsequently, the same set-up is used for the purpose inference models (paragraph 6.3), where the model estimation for the purpose inference is principally different from the zonal allocation. Next, the identification of concession-traversing transfers is discussed (paragraph 6.4). Finally, the conclusions from the model estimations are listed (paragraph 6.5).

6.1 Rule based reference models

The goal of constructing reference matrices is to investigate the effect of more complex models and evaluate if the extra effort is worthwhile. Since there is no ground truth to compare the OV-chipkaart data with, the reference matrices provide an alternative assessment option. The reference models are based on rule-based processing and therefore easy to apply. The literature provides several reference studies that apply similar procedures which attain reasonable results (see Chapter 2).

6.1.1 Zonal allocation

Origins and destinations can be directly allocated to trips based on the zones the used stops are situated in. That is, the origin of a trip is the zone where the first boarding stop in situated in, and the destination of the trip is the zone where the last alighting stop is situated in. This is done by matching the used stops in the trips dataset to their coordinates and successively allocating origins and destinations to, respectively, the first and last used stops in the GIS software of MapInfo Professional.

The travel analysis on WROOV data shows approximately half of the origins and destinations are situated in the same zone as the used stop, hence substantial discrepancies occur when using individual allocations to zones based on the zone where the used stop is situated. The cause of this large discrepancy becomes clear when the stop locations are depicted in the VENOM zonal grid (Figure 27). Many stops are located on arterial roads of the city. These roads frequently form the borders of the zonal grid.

Currently, rule-based processing is commonly applied in applications of smart card data. Therefore, we applied this simple method as reference for the improvement of an probabilistic approach.

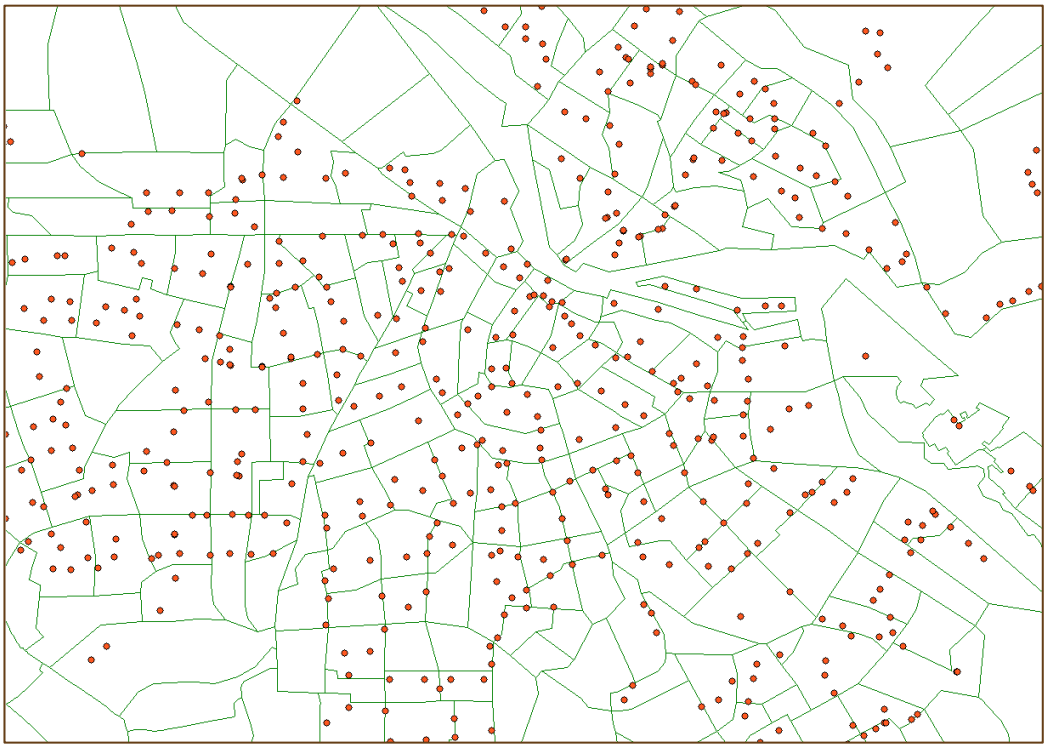


Figure 27: Stop locations in the VENOM zonal grid of Amsterdam

6.1.2 Purpose inference

The reference model for purpose inference is based on Rule-based processing, since the aim is to construct a simple model for comparison. Inspired by the study of Chakirov and Erath, we analysed the travel purpose distribution over the activity duration. This resulted in the classification of purposes based on three time intervals:

- The purpose *work* is inferred for tours with activities longer than 8 hours;
- The purpose *education* is inferred for tours with activities between 6 and 8 hours;
- The purpose *other* is inferred for tours with activities shorter than 6 hours.

The purpose *shopping* is not distinguishable from *other* by the activity duration. Accordingly, Chakirov and Erath did not incorporate this purpose. In order to construct comparable matrices, we allocated one fourth of the trips with the purpose *other* to the purpose *shopping*. This share is based on the overall distribution of purposes and is distributed randomly.

Non-home-based trips, which lack an activity duration, are allocated to the purpose other, since this purpose makes up the largest part of non-home-based trips. Travel behaviour analysis shows these rules are crude simplifications. However, the overall shares of purposes seem to match reasonably well.

6.2 Probabilistic zonal allocation models

The zonal allocation is performed by means of a multinomial logit model, estimated based on WROOV data with the Biogeme software (Bierlaire, 2003). The theoretical framework of multinomial logit modes can be found in paragraph 3.8.1. The procedure allocates trip ends, which consist of the used stops at either end of the trip, to the traffic analysis zones, of which the OD matrices are composed. The model estimates

the chance of each available zone to be "chosen", based on characteristics of that zone. In this case, the zones situated near the used stop comprise the available alternatives, also referred to as the choice set. As discussed in paragraph 3.8.2, this problem does not actually consist of a choice problem, but an allocation problem. However, throughout this chapter the terminology of discrete choice modelling is used.

Two different approaches concerning the classification of trip ends have been applied during the estimation process: trip-based and tour-based. In the trip-based approach, trip ends are considered separately for every trip. That is, for every trip the origin zone is allocated to the first boarding stop and the destination zone is allocated to the last alighting stop. The end-result consists of a trip from the origin zone to the destination zone. In the tour-based approach, on the other hand, trips are considered as linked within tours (see paragraph 1.4 for the definitions). In this context, the trip ends are classified as home-end and activity-end. The first trip of the day is assumed to start at home, hence the home zone is allocated to the first boarding stop, and the activity zone is allocated to the last alighting stop. Then, the return trip is allocated to the same activity zone and home zone, which creates consistency between trips in tours.

6.2.1 Alternative generation

The choice set generation embodies the identification of relevant alternatives, in this case zones at trip-ends. We assume that origins and destinations are in the vicinity of the used stops. Hence, the identification of alternative zones is based on the catchment areas of stops. There are, however, no definite boundaries of catchment areas. From the travel analysis on characteristics related to access and egress trip legs (see paragraph 5.1), we concluded that substantial differences exist between the sizes of catchment areas of specific stops. Although we have shown that the distances vary significantly between modes, and not between trip end classifications, we were not able to allocate the full variation of access and egress distances to specific attributes that are available in the WROOV data. Therefore, different approaches in the estimation of catchment areas have been investigated. In order to keep the procedure of choice set generation by means of catchment areas manageable, we limited the shape of catchment areas to spherical. The approaches vary in the radius of the catchment areas:

1. Uniform catchment areas: an equal radius of 400 metres for all stops, with the constraint that zones contain at least 1% of catchment area. This distance is based on concession requirement of a stop within 400 metres for all residents in Amsterdam;
2. Mode-specific catchment areas: a radius of 750 metres for metro stations, 600 metres for tram stops and 500 metres for bus stops. These distances are based on the 75 percentile of the distance covered during access and egress trip legs per mode. For simplicity reasons, specific home-end and activity-end catchment areas have not been applied, as these would require substantial added handling times;
3. Stop-specific catchment areas: Individual radius per stop, based on 90 percentile distances covered during access and egress trip legs. The 90 percentile is only applied to stops with more than ten observations, with a minimum of 500 metres and a maximum of 2500 metres. A radius of 500 metres is applied to stops with less than ten observations.
2.

The zones that overlap with the catchment areas are identified as alternative. The number of alternatives, then, varies, depending on the size of the catchment areas and to the location of the stop relative to zonal borders. For example, a stop at the centre of a large zone might just yield one alternative and a stop with a large catchment area in a zonal grid with small zones might yield many alternatives.

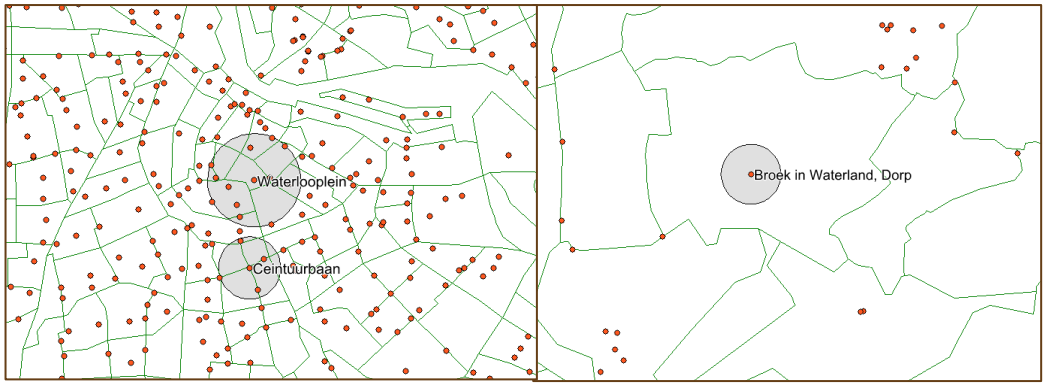


Figure 28: Catchment areas of stops

Maximum number of available alternatives determines the number of zones to be defined in the model specification. This number depends on the radius of catchment areas. Table 3 provides the maximum number of alternatives for the three different approaches of choice set generation.

Table 13: Characteristics of different choice set generation amplifications

Catchment areas	*maximum radius (m)*	*maximum number of zones in catchment area*
Uniform	400 (all)	12
Mode-specific	750 (metro)	17
Stop-specific	2500 (upper bound)	54

As indicated in the previous paragraph, trips with origins and destination outside the catchment area of the used stop are filtered from the dataset. This means that the different classifications of trip ends result in different sizes of the final datasets used for the model estimation. Table 14 provides the shares of trip-ends within the catchment areas of the used stop for the three approaches of catchment areas and both the trip-based and the tour-based classification of trip-ends.

Table 14: Shares of trip-ends within catchment areas

Catchment areas	*Trip-based approach*		*Tour-based approach*			
	All rips		Tours		Non-home-based trips	
	Origin	Destination	Home	Activity	Origin	Destination
Complete dataset	204 041 (100%)	204 041 (100%)	91 662 (100%)	91 662 (100%)	20 709 (100%)	20 709 (100%)
Uniform	148 835 (72.9%)	148 655 (72.9%)	63 294 (69.1%)	70 091 (76.5%)	15 451 (74.6%)	15 274 (73.8%)
Mode-specific	159 634	159 516	67 715	75 468	16 450	16 338

	(78.2%)	(78.2%)	(73.9%)	(82.3%)	(79.4%)	(78.9%)
Stop-specific	169 535	169 955	71 059	77 364	17 369	17 591
	(83.1%)	(83.3%)	(77.5%)	(84.3%)	(83.9%)	(84.9%)

The WROOV dataset contains only trip legs within the structural boundaries of the concessions Amsterdam and Waterland, analogous to the OV-chipkaart dataset available for this study. Hence, the fact that origins and destinations are not within the catchment areas of the used stops could indicate a transfer to another system. At the tour-based level, the difference between home-end and activity-end, which does not exist at the trip-based level, implies that there is a smaller share of transfers to other systems at the activity-end compared to the home-end. This finding suggests that Amsterdam generates more inbound activities than outbound activities. Considering the high attraction of the capital, this seems a consistent finding. However, the share of trip-ends within the catchment area also increases with larger radii of catchment areas. This implies that a small share of travellers covers a larger distance during access or egress transport.

6.2.2 Available attributes and expectations

The available attributes for the estimation of the trip-end zone consist of zonal characteristics. Most of these characteristics are obtained from the VENOM data, which have the same format as the zonal data from the LMS/NRM models. Several attributes have been created from VENOM data, WROOV data or both, or were constructed with the MapInfo GIS software (see Table 15):

- The attribute *share of catchment area* [%] consists of the areal share of the catchment area that overlaps with the respective zone;
- The attribute *distance to centroid* [m] encompasses the Euclidean distance from the used stop to the respective zonal centroid, where the zonal centroid represents the gravitational centre of the zone;
- The attribute *schools* [student places] consist of the sum of student places at secondary schools and vocational education. The student places of primary school are not considered relevant, as these students generally do not travel with PT. Moreover, the student places at higher education are not considered because these student have student PT cards (SOV) which are not included in the target population of the WROOV surveys;
- The *household size* [residents/household] is the average number of residents per household;
- *The level of urbanization* [addresses/ha] is the number of households situated in a zone divided by the zonal area;
- The attribute *cars* [cars/household] consists of the average number of cars available per household;
- The *stop density* [stops/ha] is the number of public transport stops situated in a zone divided by the zonal area;

The available attributes for the estimation of the origin zone are categorized into (1) attributes regarding the geographic information of the zones, (2) attributes regarding the build environment in zones and (3) attributes regarding the population within zones. These categories are applied in the model optimization strategy, which is presented in the next paragraph.

Table 15: Available attributes for zonal allocation and their expected effect on utility

Category	*Attribute*	*Source*	*Origin*	*Destination*	*Home*	*Activity*
Geographical	Share of catchment area	Created with VENOM	++	++	++	++
	Distance to centroid	Created with VENOM	--	--	--	--
	Area	VENOM	+	+	+	+
Activity end	PC6	postal codes	+	+	+	+
	Households	VENOM	+	+	++	+
	Urbanization	Created with VENOM	++	++	++	++
	Stops	WROOV	+	+	0	0
	Stop density	Created with stops and VENOM	+	+	+	++
	Jobs	VENOM	+	+	0	++
	Schools	Created with VENOM	+	+	0	+
Home end	Residents	VENOM	+	+	++	+
	Household size	Created with VENOM	0	0	+	-
	Working residents	VENOM	+	+	++	0
	Students	VENOM	+	+	+	0
	Cars	Created with VENOM	-	-	-	0
	Average income	VENOM	-	-	-	0

The expectations of the attributes are based on factors that influence the production and attraction in trip generation modelling. In the trip-based approach, production and attraction cannot be considered separately because the dataset consists of both away trips and return trips. The tour-based approach does provide the opportunity to distinguish differences between factors related to trip production and factors related to trip attraction, which is displayed in Table 15.

Evidently, some of the attributes are highly correlated, which should be avoided in the estimation of logit models. Highly correlated attributes in the model impair the interpretation of the influence that specific attributes have on the utility. Where other attributes were initially removed, these attributes have been investigated to compare their influence. This holds for the number of postal codes, the number of households and the number of residents. These three factors indicate the amount of potential travellers at different levels of resolution: at the level of housing blocks, houses and individuals. In addition, the geographical attributes *share of catchment area* and *distance to centroid* are also correlated: the share of the catchment area will increase as the distance to the zonal centroid decreases. It was expected that at high resolution, the share of catchment area would have more explanatory value compared to the distance to the zonal centroid since the actual trip-end might be close to the zonal border. With a wider scope, however, the explanatory value of the distance to the centroid was expected to increase, since the intra-zonal distances become

relatively smaller. Larger catchment areas might contain entire zones, so the share of catchment area is determined by the zonal area.

The tables of correlations in zonal data and trip data can be found in Appendix C.

6.2.3 Model enhancement

The model enhancement strategy involved the selection of attributes based on the interpretation of the following values:

- Significance
- Parameter value
- Rho squared statistic
- Correlation of attributes

The evaluation of individual parameter values is based on t-tests, which assess if the parameter value is significantly different from zero at a certain confidence level and a number of degrees of freedom. The degrees of freedom are determined by the sample size, which is large in this case. Therefore, the t-test values are assessed at infinite degrees of freedom. At a 95% confidence level, a t-value larger than 1.96 indicates a significant difference between the parameter value and zero. At the 99% confidence interval, the t-value should be larger than 2.575 to indicate a significant difference and at a 99.9% confidence interval this is 3.291. Because the available dataset is relatively large, t-test values are relatively high, which may cause the significance of parameter values to be overestimated. Therefore, additional assessment criteria are required to evaluate the implementation of specific attributes in the model.

The sign of the parameter value indicates if the attribute has a positive or negative effect on the utility, and thus on the probability an alternative is chosen. Parameter values with an influence that cannot be theoretically substantiated are considered undesirable, as these parameters might result in overestimation of the model on the dataset. With the size of the available data set, this is an apparent issue.

High correlations between attributes are undesired since including both attributes results in multicollinearity. The individual effects of both parameters cannot be interpreted when they partly describe the same effect. Logit models do allow little multicollinearity, but in case of two highly correlated attributes it is preferable to include only one in the model.

The goodness of fit of the model can be evaluated by means of the Rho-squared statistic, also referred to as the likelihood ratio index. This statistic specifies the increase in log-likelihood of the estimated model compared to the null-model, where the log-likelihood indicates the probability that the dataset is described by the estimated model. The value of the Rho-squared statistic lies on the interval from 0 to 1. A value of zero indicates no improvement compared to the null model, and a value of one indicates a perfect fit, which is practically impossible in logit modelling. The likelihood-ratio test provides the possibility to compare different model configurations and asses if one configuration is significantly better than the other configuration.

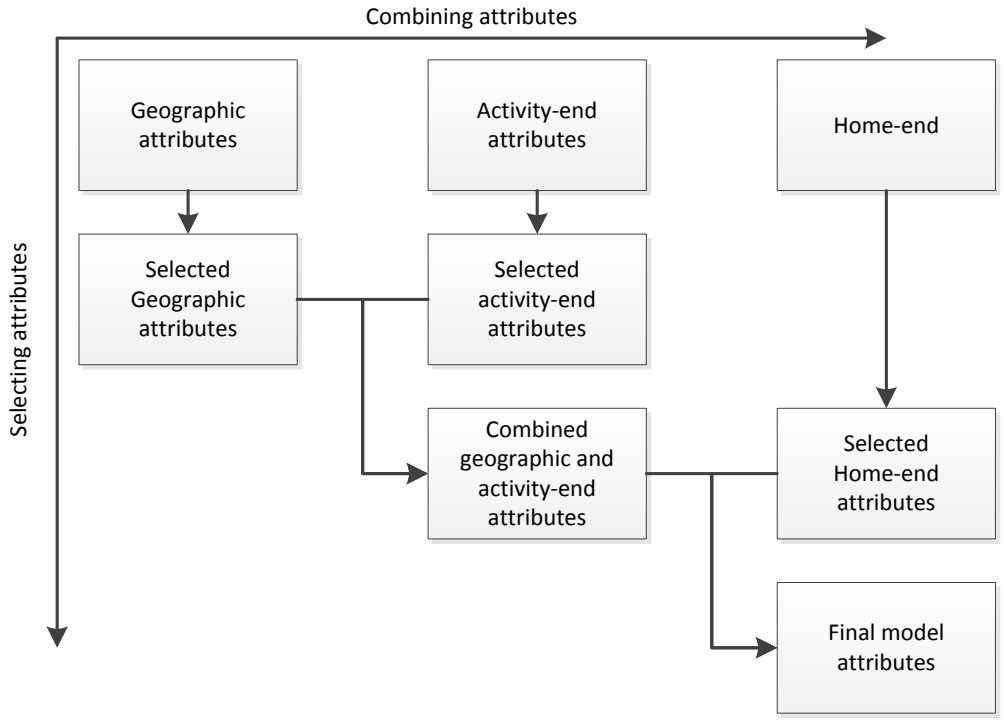
The formula of the Rho-squared statistic (Train, 2009):

$$\rho^2 = 1 - \frac{LL(\hat{\beta})}{LL(0)} \qquad 6.1$$

Where ρ^2 is the Rho-squared statistic, $LL(\hat{\beta})$ is the log-likelihood of the estimated model with parameters $\hat{\beta}$ and $LL(0)$ is the log-likelihood of the model with all parameters set to zero.

With this many available attributes, the number of possible model configurations is too large to assess all of them. In order to identify viable model configurations, a sequential strategy has been applied, with the following steps:

1. A model with only alternative specific constant parameters was estimated to test if the randomization resulted in generic alternatives, independent from the alternative number. All alternative specific constants tested insignificant, verifying that the utilities of alternatives are independent of the alternative numbering;
2. The different categories of attributes were assessed separately. This provided insight in the explanatory value of each attribute block and filtered irrelevant and highly correlated parameters. The optimization of each category block resulted in a selection of attributes with explanatory value, that are significant and have minimal correlation;
3. The remaining attributes of different category blocks are combined. Because the estimated parameters in logit models are all relative, combining attribute blocks results in additional correlations that influence the model stability. Therefore, after adding an addition block, another selection sequence has been applied.
3.



4.

Figure 29: Model enhancement strategy for zonal allocation

6.2.4 Convergence complications

During the model optimization process, we encountered complications regarding the model convergence. Depending on the applied algorithm, Biogeme then produces output with one of the following the diagnostics:

- Radius of the trust region is too small (BIO algorithm)
- No significant improvement possible (DONLP2 algorithm)

- Normal termination (SOLVOPT algorithm)

In case the model has not converged, the estimated parameter values cannot be interpreted as valid estimates. Therefore, the cause of these complications has been investigated.

"*Radius of the trust region too small*" can indicate that the model is close to singular which can be caused by inclusion of irrelevant parameters or totally correlated parameters. However, in the majority of model configurations the smallest singular value was found to be close to 200. This leads to believe that the converging issues were not caused by singularity of the matrix.

When confronted with convergence complications, Bierlaire recommends assessment of the final gradient norm, which should be close to zero. The final gradient norm differs between optimization algorithms, but the estimated parameters are equal. In some cases the DONLP2 algorithm diagnoses convergence, where the BIO algorithm does not. The SOLVOPT algorithm generally estimates a larger final gradient norm and converges less easily.

Convergence complications are mainly caused by correlations between attributes. The issues caused by correlated attributes can be solved by excluding one of the two correlated parameters. This exclusion is at the expense of the model fit. However, in several instances, a model with only the attribute distance to the centroid still resulted in an unstable model. Hence, correlations between attributes were not the only cause of model instability. We have not been able to discover other causes for instability, but these might be related to the high values of land-use characteristics in relation to the low variance between adjacent zones. However, this theory has not been investigated.

6.2.5 *Final parameter estimates*

The following tables present the final model statistics and parameter estimates for the trip-based model (Table 16), the tour-based model (Table 17) and the model for non-home-based trips (Table 18). Only the final model configurations are presented here, intermediate results from the estimation process are presented in Appendix C.

The final model configurations only include two or three attributes, since most available attributes were insignificant or caused instability of the model. Paragraph 6.2.7 deliberates on the interpretation of the individual model parameters.

Table 16: Final estimation results of trip-based zonal allocation models

Model statistics	*Origin Inference*		*Destination Inference*	
Model	Multinomial Logit		Multinomial Logit	
Number of parameters	3		3	
Rho-square	0.322		0.319	
Diagnostic	Convergence reached...		Convergence reached...	
Parameters	Value	t-test	Value	t-test
$\beta_{i,1}$ (share of catchment area)	0.0851	320.46	0.0851	320.82
$\beta_{i,2}$ (stop density)	2.97	170.15	2.94	170.22

$\beta_{i,3}$ (urbanization level)	0.00183	16.78	0.00143	13.14

The estimated trip-based zonal allocation models are represented in formula by the following equations. Equation 6.3 presents the *origin zone allocation model* and equation 6.4 presents the *destination zone allocation model*. The model parameters and the model fit are approximately similar. This can be explained by the trip-based definition of trip-ends in combination with the high share of symmetry in the trip-based data set used for estimated. As a result, access legs in the away trip are made in opposite direction during the return trip as egress legs. The slight differences between origin and destination allocation can be attributed to the differences in non-home-based trips.

$$V_{O,i} = 8.51 * 10^{-2} * X_{i1}\,[\%] + 2.97 * X_{i2}\,[stops/ha] + 1.83 * 10^{-3} * X_{i3}\,[addresses/ha] \qquad 6.2$$

$$V_{D,i} = 8.51 * 10^{-2} * X_{i1}\,[\%] + 2.94 * X_{i2}\,[stops/ha] + 1.43 * 10^{-3} * X_{i3}\,[addresses/ha] \qquad 6.3$$

With:
$V_{O,i}$ = systematic utility of alternative *i* for origin zone *O*
$V_{D,i}$ = systematic utility of alternative *i* for destination zone *D*
X_{i1} = value of share of catchment area of alternative *i*
X_{i2} = value of stop density of alternative *i*
X_{i3} = value of urbanization level of alternative *i*

Table 17: Final estimation results of tour-based zonal allocation models

Model statistics	*Home zone inference*		*Activity zone inference*	
Model	Multinomial Logit		Multinomial Logit	
Number of parameters	3		2	
Rho-square	0.347		0.311	
Diagnostic	Convergence reached...		Convergence reached...	
Parameters	Value	t-test	Value	t-test
$\beta_{i,1}$ (share of catchment area)	0.0824	81.69	0.0901	92.56
$\beta_{i,2}$ (stop density)	2.81	36.78	2.86	51
$\beta_{i,3}$ (urbanization level)	0.00613	14.23	0	(fixed)

The estimated tour-based zonal allocation models are represented in formula by the following equations. Equation 6.5 presents the *home zone allocation model* and equation 6.6 presents the *activity zone allocation model*. In contrast to the trip-based models, the tour-based zonal allocation models differ substantially between trip-ends. At the activity end, the influence of the urbanization level did not prove to be significant and stable and therefore was omitted from the model. At the home end, the same attributes are included as in the trip-based models. The fit statistics indicate a better fit of the home zone allocation model compared to the activity zone allocation model, which can be explained by the additional attribute.

$$V_{H,i} = 8.24 * 10^{-2} * X_{i1}\,[\%] + 2.81 * X_{i2}\,[stops/ha] + 6.13 * 10^{-3} * X_{i3}\,[addresses/ha] \qquad 6.4$$

$$V_{A,i} = 9.01 * 10^{-2} * X_{i1}\,[\%] + 2.86 * X_{i2}\,[stops/ha] \qquad 6.5$$

With:
$V_{H,i}$ = systematic utility of alternative *i* for home zone *H*
$V_{A,i}$ = systematic utility of alternative *i* for activity zone *A*
X_{i1} = value of share of catchment area of alternative *i*
X_{i2} = value of stop density of alternative *i*
X_{i3} = value of urbanization level of alternative *i*

In order to complete the tour-based modelling approach, also non-home-based trips have to be considered. These trips are not part of a tour, so the tour-based zonal allocation models are not applicable. Moreover, non-home-based trips have different travel patterns than home-based trips (see appendix B). Therefore, specific non-home-based zonal allocation models are required.

However, the quantitative comparison of the data sources has shown that the share of non-home-based trips is substantially larger in the OV-chipkaart dataset compared to the WROOV dataset. Consequently, the WROOV data are not optimal for estimation of the non-home-based models. Since no better alternatives are available, we continued with the WROOV data.

Table 18: Final estimation results of non-home-based trip zonal allocation models

Model statistics	*Origin inference*		*Destination inference*	
Model	Multinomial Logit		Multinomial Logit	
Number of estimated parameters	3		3	
Number of observations	2631		2628	
Rho-square	0.377		0.345	
Diagnostic	Convergence reached...		Convergence reached...	
Parameters	Value	t-test	Value	t-test
$\beta_{i,1}$ (share of catchment area)	0.0519	17.9	0.0525	18.53
$\beta_{i,2}$ (stop density)	2.55	18.77	2.33	18.71
$\beta_{i,3}$ (distance to centroid)	-0.00222	-14.75	-0.00219	-15.03

The estimated non home-based zonal allocation models are represented in formula by the following equations. Equation 6.7 presents the *origin zone allocation model* and equation 6.8 presents the *destination zone allocation model*.

$$V_{O,i} = 5.19 * 10^{-2} * X_{i1}\,[\%] + 2.55 * X_{i2}\,[stops/ha] - 2.22 * 10^{-3} * X_{i3}\,[m] \qquad 6.6$$

$$V_{D,i} = 5.25 * 10^{-2} * X_{i1}\,[\%] + 2.33 * X_{i2}\,[stops/ha] - 2.19 * 10^{-3} * X_{i3}\,[m] \qquad 6.7$$

With:
$V_{O,i}$ = systematic utility of alternative *i* for origin zone *O*
$V_{D,i}$ = systematic utility of alternative *i* for destination zone *D*
X_{i1} = value of share of catchment area of alternative *i*
X_{i2} = value of stop density of alternative *i*
X_{i3} = value of distance to centroid of alternative *i*

6.2.6 Model stability analysis

In order to assess the transferability of the models, a stability analysis has been performed on the final model configurations. This analysis comprises the estimation of these models on the yearly WROOV data sets, for the years 2003 to 2009.

From this analysis it can be concluded that the parameter estimates for the attributes *share of catchment area* and *stop density* are stable over the years. The parameter estimate for the *urbanization level* is less stable and shows a larger variation over the years. This was also indicated by the relatively low t-value compared to the other parameters. The distinction between home-end and activity-end shows that this variation is mainly caused by the urbanization at the activity-end. For the activity allocation model, this attribute is excluded because it unsettles the model convergence. In the home allocation models, on the other end, the parameter estimates are more stable over time compared to the trip based approach.

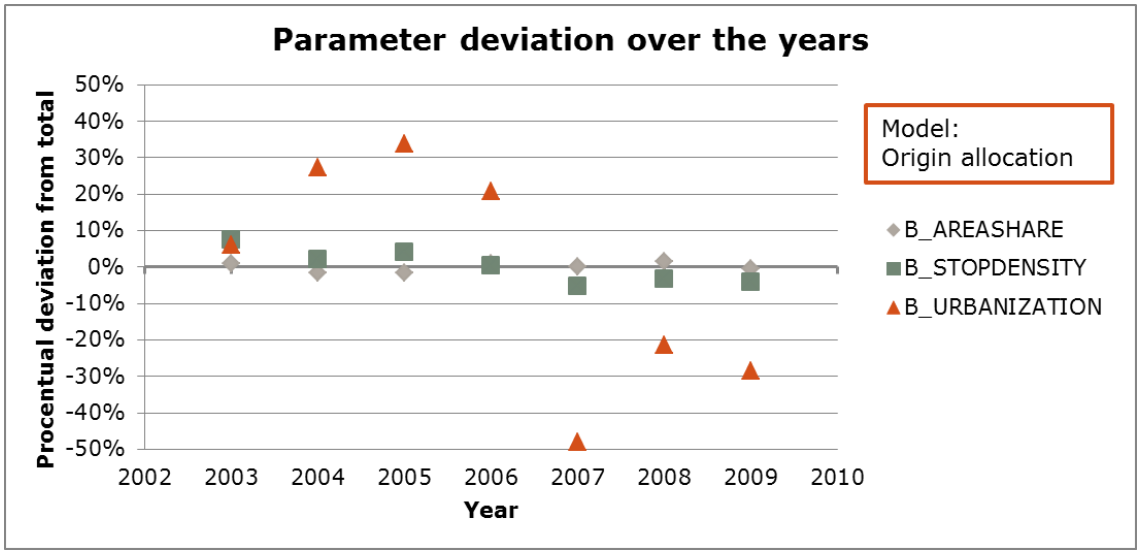


Figure 30: Origin allocation model parameter values for yearly WROOV datasets

The relative deviation of the coefficient for the urbanization level in the trip-based models shows a distinction between two periods: up to 2006 and after 2006 (see Figure 30). The effect of the urbanization level is estimated higher in the years up to 2006, compared to the effect estimated on the entire dataset, and lower in the years after. This is counterintuitive since the urbanization level is determined from data of the year 2010. The distinction between these periods is not observed for the home zone allocation model (see Figure 31). The coefficient for urbanization level still deviates more than the coefficients for share of the catchment area and stop density, but substantially less compared to the deviation in the origin allocation model.

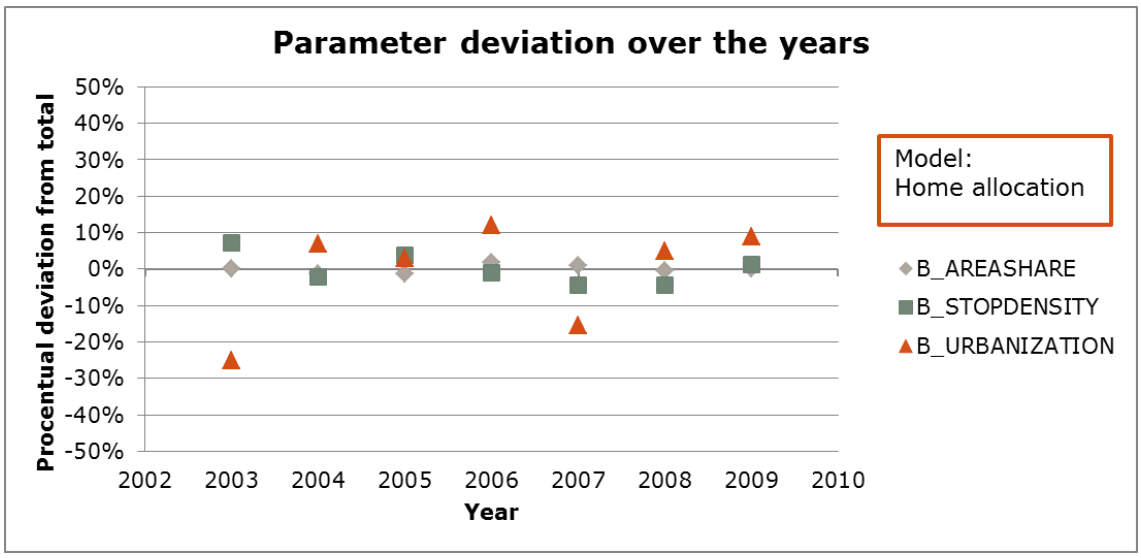


Figure 31: Home allocation model parameter values for yearly WROOV datasets

6.2.7 Interpretation of the model parameters

The trip-based models for origin allocation and destination allocation consist of almost equal parameters, which was expected due to the mix of away trips and return trips in the dataset. First, we interpret the three attributes in the trip-based zonal allocation models, before comparing the coefficients with the distinct tour-based zonal allocation models.

First, the geographical attribute share of the catchment area has the largest effect on the model fit, based on the Rho-squared increase) and proved very stable over time. A higher share of the catchment area in a zone results in a higher utility for that zone, and thus a higher probability. This large effect was expected, since travellers tend to optimize their travel time. Especially in a highly urban environment like Amsterdam, the public transport supply allows travellers to choose stops situated near their origins and destinations, resulting in short access and egress legs. Consequently, origins and destinations are more likely to be near the used stop.

Second, a higher stop density also results in a higher probability of that zone being the origin or destination zone. Since stop locations are deliberately positioned near locations with high public transport demand, these can be interpreted as indicators of trip production and attraction. The stability analysis also indicates a stable effect of the stop density on the probabilities of zones over time.

Third, and last, the level of urbanization indicates a positive relation with the utility of zones. Similar to the stop density, the density of addresses in a zone relates to the trip production and attraction. Addresses are potential origins and destinations, explaining the higher probability of zones with a higher urbanization level.

Compared to the trip-based approach, the tour-based zonal allocation models show differences between the trip-ends. Most notably, the level of urbanization is not included in the activity zone allocation model, since it has a small negative effect on the utility. Moreover, the effect was not significant for all years. Therefore, the attribute was excluded from the activity allocation model. At the home end, the coefficient of the level of urbanization more than tripled compared to its value in the

origin allocation model. Since the coefficients for the share of the catchment area and the stop density are very similar to the coefficients in the origin allocation model, the influence of the level of urbanization on the utility of alternative zones is also greater. This indicates that the effect of the urbanization level in the trip-based zonal allocation models is in fact an averaged effect of the large influence at the home-end and the absent effect at the activity end.

The fact that, of all available land-use attributes, only the stop density and the urbanization are included in the final model results from insignificant or unstable effects of other land-use characteristics. Especially for the activity zone allocation, we expected to find significant effects for the number of jobs and schools in a zone. While these attributes showed significant effects before combining them with geographic attributes, the combination destabilized the model, as it did not converge. A model configuration with only the attributes share of the catchment area and number of jobs did converge. However, replacing the number of jobs by the stop density resulted in a significantly better model fit.

The stop density does not directly relate to origins and destinations of trips, as stops are not the actual locations where the distinguished activities are performed. Hence, the stop density serves as a proxy for activity locations. Since stops are generally located near activity sites, the stop density can be interpreted as an indirect indicator of activity locations. It is preferable to include only direct indicators in the model, since the interpretation of indirect attributes cannot be completely disconnected from other factors. Moreover, direct attributes are more stable over time. Nonetheless, the stop density has been included in the absence of direct indicators of activity locations.

The attributes *share of the catchment area* and *distance to the centroid* are highly correlated. By including both attributes, the model stability was compromised and the interpretation of individual attributes distorted by multicollinearity. Therefore, we have only included one of them in the trip-based and tour-based zonal allocation models. We have chosen the attribute share of the catchment area over the distance to the centroid for three reasons. The primary reason is the better model fit of the share of the catchment area. Second, the attribute distance to the centroid resulted in instability of the models. In several instances, the model did not converge with only this attribute. We have not found the exact cause of this instability, but it might be that the relation of the distance to the centroid is not linear with the log odds of the alternatives. Third, the share of the catchment area is considered a better indicator of the nearness of a zone to the used stop than the distance to the centroid in case of small catchment areas. The distance of the centroid depends on the size of the zone and its relative location to the used stop. In case larger catchment areas are applied, the distance to the centroid showed a better fit than the share of the catchment area.

In the non-home-based zonal allocation models, we aimed for the highest model fit. Since the available dataset of non-home-based trips from WROOV is not comparable to the non-home-based trips in OV-chipkaart data we focussed on the tour-based models. In order to limit the effects of non-home-based trips, the interpretation of the models was considered less important. The models include the correlated attributes share of the catchment area and the distance to the centroid, since including both did not compromise the model stability. As a result, the interpretation of the individual attributes is distorted. However, it can be concluded that the effects of share of the catchment area and stop density are similar to the effects in the trip-based and tour-based models. The distance to the centroid has a negative effect on the utility. This

was expected, as travellers tend to optimize their travel time, and thus minimize the access and egress distances.

In order to get a feeling of the effect of the attributes on the probabilities in the different models, we present two examples of trip-ends with corresponding alternatives and their probabilities: example 1 in Table 20 and example 2 in Table 21. The values in the examples are fictional, but based on the descriptive statistics from the dataset presented in Table 19.

Table 19: Descriptive statistics of the attributes in the zonal allocation models

Attributes	*Minimum*	*Maximum*	*Mean*	*Std. Deviation*
Share of catchment area [%]	0	100	14.1	15.7
Stop density [stops/ha]	0	1.25	0.18	0.15
Level of urbanization [addresses/ha]	0	207	48.9	41.8

The first example shows that, with few alternatives, the influence of the attribute share of the catchment area is overestimated due to the relative scale, which results in high absolute differences of the percentage between zones. With a higher stop density and higher urbanization level, zone 2 was expected to obtain a higher probability than 4%. The relative scale of the share of the catchment area results in heteroscedasticity of the residuals. The probabilities differ per approach, caused by the different effects of the level of urbanization. Even though the differences are small, it does clearly indicate the absence of the urbanization level in the activity allocation and the averaging effect in the trip-based models.

Table 20: Zonal allocation example 1

Attributes	*Zone 1*	*Zone 2*
Share of catchment area [%]	80	20
Stop density [stops/ha]	0.1	0.5
Urbanization level [addresses/ha]	100	200
Probabilities	$P_{n,1}$	$P_{n,2}$
Origin zone allocation	98%	2%
Destination zone allocation	98%	2%
Home zone allocation	96%	4%
Activity zone allocation	99%	1%

Example 2 is more representative for the dataset, as it has more alternatives. Consequently, the differences between the values of share of the catchment area are smaller. Consequently, the probabilities are more dispersed over the alternatives. However, the effect is still very large, which complies with the effect on the model fit. The effect of the stop density is also substantial, looking at the probabilities of alternatives 4 and 5. The effect of the urbanization level is not prominent in the trip-based models. On the other hand, the tour-based models do show a larger effect. In the home zone allocation model, differences are noticeable between the probabilities

of zones 2 and 3, while the probabilities of these zones are equal in the activity zone allocation.

Table 21: Zonal allocation example 2

Attributes	*Zone 1*	*Zone 2*	*Zone 3*	*Zone 4*	*Zone 5*	*Zone 6*
Share of catchment area [%]	5	10	10	20	20	35
Stop density [stops/ha]	1	0.2	0.2	0.2	0.5	0.2
Level of urbanization [addresses/ha]	150	100	200	100	100	100
Probabilities	$P_{n,1}$	$P_{n,2}$	$P_{n,3}$	$P_{n,4}$	$P_{n,5}$	$P_{n,6}$
Origin zone allocation	29%	4%	5%	9%	22%	32%
Destination zone allocation	28%	4%	4%	9%	22%	32%
Home zone allocation	32%	4%	7%	9%	20%	29%
Activity zone allocation	24%	4%	4%	9%	22%	36%

6.3 Probabilistic purpose inference models

The purpose inference models have been approached separately from the zonal allocation models. The model structure is different due to the pre-determined number of alternatives, which are not generic. Because the influence of trip characteristics is not equal for the distinct purposes, these can be implemented as model attributes (see also paragraph 3.8.6).

6.3.1 Alternative selection

Although the WROOV study specified more alternatives, only four alternatives are distinguished in this study. The purposes *work*, *education* and *shopping* are the most frequent travel purposes in bus and light rail. Remaining purposes are aggregated in the purpose *other*.

The purposes work, education and shopping are the most relevant purposes for long term forecasts. Besides the fact that they are the most frequent purposes, they are also the most susceptive for policy measures. Commuting traffic is influenced by policy on work hours, flex-workers, home-workers. Educational traffic is influenced by policy measures regarding the student PT cards and study financing. Shopping traffic is influenced by increased opening hours of shops at specific locations. On the other hand, visiting family and friends or hospitals are less prone to policy measures due to their optional character. The identification of business trips would be interesting regarding the influence of policies, but it generates few trips with bus and light rail and therefore cannot be distinguishes from other discretionary purposes.

The travel data analysis has indicated that the compulsory purposes *work* and *education* are correlated in terms of key variables. The same holds for the discretionary purposes *shopping* and *other* (see paragraph 5.2.1). Nonetheless, we have aimed at the inference of these four purposes due to the inclusion of the activity duration and contract duration.

6.3.2 Available attributes and model structure

The logit models estimate relative utilities of the alternatives. Therefore, one alternative has to be normalized. Since only differences in the utilities matter to the final probabilities of alternatives being chosen, the most logical normalization is to set one alternative utility to zero. Because the purpose other is the least specific of the distinguished travel purposes, we have normalized this purpose to zero.

The categorical attributes are implemented by means of dummy variables. This means for every category of an attribute a specific coefficient is estimated, with exception of one reference category. Similarly to the alternatives, the categories have to be normalized by setting one category to zero. The choices of references categories are based on our perception of the least specific category of every attribute, related to the specified purposes work, education and shopping. Hence, we aimed at the highest possible coefficients.

The available attributes, their categorization and their values are presented in

Table 22. The following attributes have been investigated during the model estimation process:

- *Concession*: the concession where the trip took place. We expected different travel patterns for (1, the reference) Amsterdam, (2) Waterland and (3) travellers traversing both concessions in their trips;
- *Mode*: the mode used during the trip, in this study limited to (1, the reference) bus, (2) tram and (3) metro or (4) multiple modes;
- *Departure time*: the departure time of the trip, categorized in five times of day: (1) the early morning, between 4 am and 7 am, (2) the morning peak, between 7 am and 9 am, (3, the reference) the day, between 9 am and 4 pm, (4) the evening peak, between 4 pm and 6 pm and (5) the night, between 6 pm and 4 am;
- *Distance travelled*: the distance covered by public transportation, based on the route travelled, measured in kilometres;
- *Number of trip legs*: the number of trip legs within the trip, also perceptible as the number of transfers made within the trip plus one;
- *Activity duration*: the duration between consecutive trips, calculated by the time between consecutive boardings, measured in minutes;
- *The contract duration*: the validity period of the travel product. Contracts durations are limited to (1, the reference) no contract, (2) year contracts and (3) month contracts;
- The frequency: the travel frequency, measured by the number of tours per week.

In addition to these trip characteristics and travel pattern characteristics, land use characteristics at both ends of the trips have been investigated. Depending on the definitions used, the trip-based origin and destination, or the tour-based home-end and activity-end, their expected influence differs. Because the exact origin and destination of trips are unknown, we aggregated the land-use characteristics over a 400 metre radius around the used stop. The analysis of access and egress distances shows that a substantial amount of access and egress legs cover longer distances than 400 metres (see paragraph 5.1). Nonetheless, we applied a conservative radius because the averaging effect over larger areas increases. This means that the land-use attributes might not reflect the exact values at the origins and destinations.

Table 22: Available attributes for purpose inference

Attribute category	*Attribute*	*Measurement level*	*Values/ [Units]*
Trip characteristics	concession	categorical	Amsterdam Waterland Both
	mode	categorical	Bus Tram Metro Multiple
	Departure time	categorical	Early morning Morning peak Midday Afternoon peak Night
	distance travelled	continuous	[km]
	number of trip legs	continuous	[trip legs within trip]
	activity duration	continuous	[minutes]
Travel pattern	contract duration	categorical	Year Month None
	frequency	continuous (-)	[trips per week]
Land use characteristics	households	continuous (-)	[addresses]
	average income	continuous (-)	[€/year]
	jobs	continuous (-)	[# of jobs]
	residents	continuous (-)	[# of residents]
	schools	continuous (-)	[# of student places]
	students	continuous (-)	[# of students]
	working population	continuous (-)	[# of working residents]

Since we expected different effects of attributes on specific purposes, we implemented purpose-specific coefficients for each attribute. In combination with the implementation of dummy variables for each categorical attribute, this resulted in a large number of parameters to be estimated.

6.3.3 Model enhancement

The model enhancement strategy applied for the purpose inference models is similar to the approach of the enhancement of zonal allocation models (see paragraph 6.2.3), but contains different categories of attributes. Figure 32 presents the framework of the selection of attributes to include in the purpose inference models.

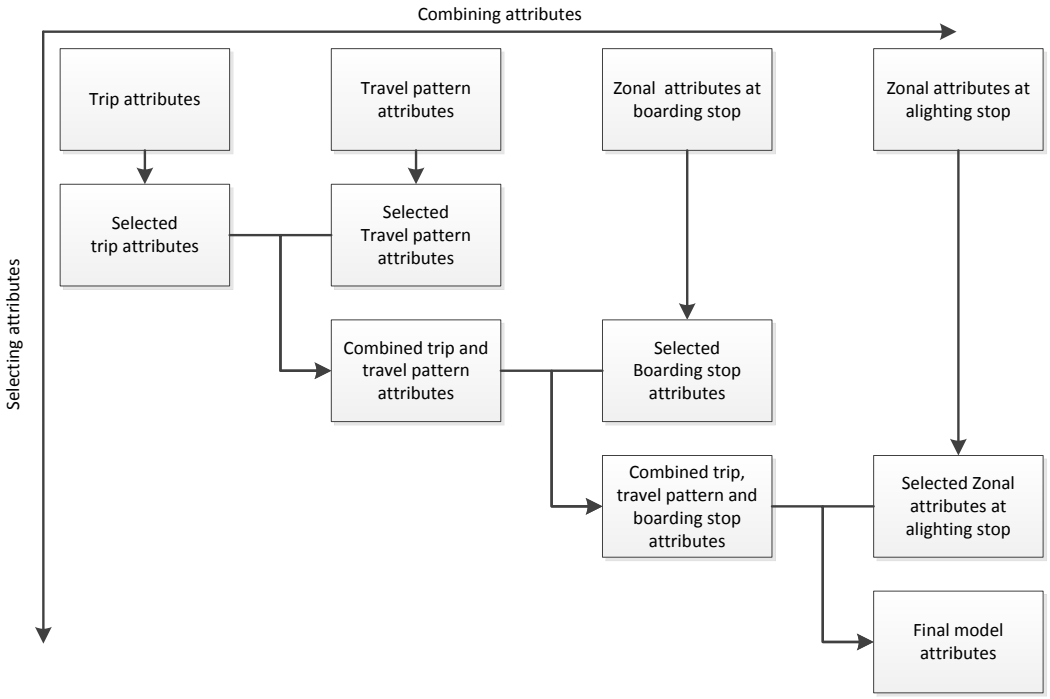


Figure 32: Enhancement strategy of purpose inference models

Since several land-use attributes are highly correlated, especially the ones related to the home end, we applied distinct attributes related to a specific purpose. At the home-end this included the number of households for the purpose shopping, the working population for the purpose work and the number of students for the purpose education. At the activity-end, the attribute *jobs* was included in the utility function for the purpose *work* and the attribute *schools* in the utility function for the purpose *education*. An attribute related to shops was not available to implement in the utility of the purpose *shopping*.

In the trip-based approach, the distinction between home-end and activity-end is not made. Therefore, all attributes were initially implemented at both trip-ends. Without alternative specific constants, the land-use attributes showed a substantial improvement of the log-likelihood. After adding alternative specific constants, however, these proved to explain the same increase in log-likelihood.

In the tour-based approach, the fit of the model configurations with only land-use characteristics was higher compared to the trip-based approach. The attributes are all significant in the configurations without trip and travel pattern attributes, but the models do not converge. The same problem with land-use attributes arose in the zonal allocation models. The exact origin of the model instability resulting from the land-use attributes has not been found.

Although the travel analysis indicated an increased share of work trips for longer distances, the travel distance was excluded from the final model configurations. Different implementation methods have been investigated: linear, quadratic and exponential, but in every case the effect was insignificant. Therefore, it is concluded that the effect of the travel distance on the utility of specific purposes is already explained by other attributes.

During the model estimation process, the same data sample was used as for the zonal allocation models. The final model attributes were than tested for stability over time on the yearly WROOV datasets. Since the longitudinal data analysis showed a slight trend in the purpose shares, the final model coefficients from 2008 are used. This was the most recent year without interference of the OV-chipkaart on the WROOV survey in the Amsterdam region.

6.3.4 Final model parameter estimates

In the final model configurations, only of trip characteristics and travel pattern characteristics are included in all model approaches. The trip-based purpose inference model incorporates five attributes, plus alternative specific constants. The tour-based approach allows for the incorporation of the additional attribute *activity duration*. Because the activity duration is correlated to several other attributes, its inclusion results in several alternations compared to the trip-based model. In the final configuration, seven attributes are included. The number of estimated parameters, however, is lower compared to the trip-based model due to the exclusion of correlated parameters. The non-home-based model only includes three attributes, as the other attributes proved insignificant.

Table 23: Model statistics for the three specific purpose inference models

Model statistics	All trips	Tours	Non-home-based trips
Model	Multinomial Logit	Multinomial Logit	Multinomial Logit
Number of parameters	25	24	16
Rho-square	0.376	0.460	0.318
Diagnostic	Convergence reached...	Convergence reached...	Convergence reached...

The model fit statistics indicate an increased fit for the tour-based model, which can be attributed to the inclusion of the attribute activity duration. Table 24 presents the estimated coefficients for each model parameter in the three model approaches. The purpose-specific coefficients make this table hard to read. It does, however, provide an easy comparison of the coefficient values in the different approaches. In order to increase the readability of the final model configurations, Table 24 is translated into formulas that describe the utility functions for each purpose in all model approaches. The utility functions are presented in equation 6.8 to equation 6.9.

Table 24: Final parameter estimates for specific purpose inference models

Parameters	*Purpose*	*All trips*		*Tours*		*Non-home-based*	
Attribute		β	t-test	β	t-test	β	t-test
ASC Education	E	-1.78	-43.64	-2.38	-35.83	-3.07	-24.07
ASC Shopping	S	-0.862	-28.99	0	(fixed)	-1.55	-19.52
ASC Work	W	-1.33	-48.26	-2.53	-27.09	-1.24	-18.47
Frequency	E	0.342	26.91	0.467	26.63	0.179	8.35
Frequency	S	-0.246	-12.81	0	(fixed)	0	(fixed)
Frequency	W	0.267	31.03	0.346	22.41	0.131	7.68
Concession Waterland	S	0.585	7.73	0.694	6.03	0	(fixed)
Concession Waterland	W	0.532	9.47	0	(fixed)	0	(fixed)
Concession Both	W	0.525	10.11	0.285	3.48	0	(fixed)
Contract duration Year	E	-0.779	-9.8	-1.61	-16.59	-1.02	-3.89
Contract duration Year	S	0	(fixed)	-0.264	-2.06	0	(fixed)
Contract duration Year	W	0.954	24.41	0	(fixed)	0.0801	0.6
Contract duration Month	E	0.849	13.75	0	(fixed)	0.534	2.5
Contract duration Month	S	0	(fixed)	0	(fixed)	0	(fixed)
Contract duration Month	W	0	(fixed)	-0.812	-11.56	-0.836	-5.17
Mode tram	E	-0.609	-12.48	-0.49	-6.62	0	(fixed)
Mode metro	S	-0.631	-9.05	-0.614	-5.98	0	(fixed)
Mode metro	W	0.662	18.53	0.527	9.06	0	(fixed)
Mode multiple	W	0.31	8.41	0.431	5.7	0	(fixed)
TOD early morning	S	0	(fixed)	-2.17	-3.68	0	(fixed)
TOD early morning	W	2.28	28.46	1.88	17.98	2.13	10.58
TOD morning peak	E	1.14	21.3	2.08	26.22	1.98	12.85
TOD morning peak	S	-1.71	-12.57	-1.52	-9.08	-1.48	-4.65
TOD morning peak	W	1.68	40.95	1.67	28.33	1.6	14.63
TOD evening peak	E	-0.79	-9.68	0	(fixed)	0	(fixed)
TOD evening peak	S	-0.464	-6.82	-0.972	-6.67	-0.629	-3.15
TOD evening peak	W	1.32	34.47	0	(fixed)	0.263	2.1
TOD night	E	-1.18	-12.09	0	(fixed)	0	(fixed)
TOD night	S	-0.999	-12.91	-0.628	-3.89	-1.17	-4.38
number of trip legs	S	0	(fixed)	-0.266	-6.4	0	(fixed)
number of trip legs	W	0	(fixed)	-0.376	-7.12	0	(fixed)
Activity duration	S			-0.00218	-10.49		
Activity duration	W			0.00466	31.33		

Trip based purpose inference model

$$V_{n,work} = -1.33 + 0.267\,X_1 + \begin{bmatrix}0.532\\0.525\end{bmatrix}\hat{X}_2 + \begin{bmatrix}0.954\\0\end{bmatrix}\hat{X}_3 + \begin{bmatrix}0\\0.662\\0.310\end{bmatrix}\hat{X}_4 + \begin{bmatrix}2.28\\1.68\\1.32\\0\end{bmatrix}\hat{X}_5 \quad (6.10)$$

$$V_{n,educ} = -1.78 + 0.324\,X_1 + \begin{bmatrix}-0.779\\0.849\end{bmatrix}\hat{X}_3 + \begin{bmatrix}-0.609\\0\\0\end{bmatrix}\hat{X}_4 + \begin{bmatrix}0\\1.14\\-0.79\\-1.18\end{bmatrix}\hat{X}_5 \quad (6.11)$$

$$V_{n,shop} = -0.862 - 0.246\,X_1 + \begin{bmatrix}0.585\\0\end{bmatrix}\hat{X}_2 + \begin{bmatrix}0\\-0.631\\0\end{bmatrix}\hat{X}_4 + \begin{bmatrix}0\\-1.71\\-0.464\\-0.999\end{bmatrix}\hat{X}_5 \quad (6.12)$$

Tour based purpose inference model

$$V_{n,work} = -2.53 + 0.346\,X_1 + \begin{bmatrix}0\\0.285\end{bmatrix}\hat{X}_2 + \begin{bmatrix}0\\-0.812\end{bmatrix}\hat{X}_3 + \begin{bmatrix}0\\0.527\\0.431\end{bmatrix}\hat{X}_4 + \begin{bmatrix}1.88\\1.67\\0\\0\end{bmatrix}\hat{X}_5 - 0.376\,X_6 + 4.66 * 10^{-3}\,X_7 \quad (6.13)$$

$$V_{n,educ} = -2.38 + 0.467\,X_1 + \begin{bmatrix}-1.61\\0.849\end{bmatrix}\hat{X}_3 + \begin{bmatrix}-0.490\\0\\0\end{bmatrix}\hat{X}_4 + \begin{bmatrix}0\\2.08\\0\\0\end{bmatrix}\hat{X}_5 \quad (6.14)$$

$$V_{n,shop} = + \begin{bmatrix}0.694\\0\end{bmatrix}\hat{X}_2 + \begin{bmatrix}-0.264\\0\end{bmatrix}\hat{X}_3 + \begin{bmatrix}0\\-0.614\\0\end{bmatrix}\hat{X}_4 + \begin{bmatrix}-2.17\\-1.52\\-0.972\\-0.602\end{bmatrix}\hat{X}_5 - 0.266\,X_6 - 2.18 * 10^{-3}\,X_7 \quad (6.15)$$

Non-home-based purpose inference model

$$V_{n,work} = -1.24 + 0.131\,X_1 + \begin{bmatrix}0.0801\\-0.836\end{bmatrix}\hat{X}_3 + \begin{bmatrix}2.13\\1.60\\0.263\\0\end{bmatrix}\hat{X}_5 \quad (6.16)$$

$$V_{n,educ} = -3.07 + 0.179\,X_1 + \begin{bmatrix}-1.02\\0.534\end{bmatrix}\hat{X}_3 + \begin{bmatrix}0\\1.98\\0\\0\end{bmatrix}\hat{X}_5 \quad (6.17)$$

$$V_{n,shop} = -1.55 + \begin{bmatrix}0\\-1.48\\-0.629\\-1.17\end{bmatrix}\hat{X}_5 \quad (6.18)$$

With: V_{ni} = systematic utility of purpose *i* for trip *n*
X_1 = value of travel frequency
$\hat{X}_2$ = unit vector of *concession* = [EBS both]
$\hat{X}_3$ = unit vector of *contract duration* = [year month]
$\hat{X}_4$ = unit vector of *mode* = [tram metro multiple]
$\hat{X}_5$ = unit vector of *departure time* = [early morning morning peak evening peak night]
X_6 = value of number of trip legs
X_7 = value of activity duration

In these utility formulas, the dummies of categorical variables are presented as unit vectors. These consist of vectors with a 1 for the represented category and zeros for

the other categories. In case of the reference category, the vector is a null vector, with only zeros.

6.3.5 Interpretation of the individual model parameters

All the estimated utilities are relative to the purpose other. Moreover, the coefficients for categorical variables are relative to the reference category. The alternative specific constants correct the estimated utilities for unobserved factors. Since these are also relative to the estimated coefficients, they do not represent a specific utility or disutility.

The travel frequency has a positive relation to the utilities of the compulsory purposes and a negative effect on the utility of shopping, although this was not significant in the tour-based model. Since the effect of the travel frequency was expected to be similar for shopping and other, this complies with the expectations.

The attribute concession shows an increased utility for the purposes work and shopping for trips made in Waterland in the trip-based model. However, in the tour-based model, trips made in Waterland do not have a higher utility for the purpose work. Trips covering both concessions do have a higher utility for the purpose work. This indicates that travellers from Waterland that commute to Amsterdam are more likely to continue their trip with GVB, while shopping travellers do not.

In the trip-based model, year contracts have a positive effect on the utility of the purpose work, while it has a negative effect for the purpose education. In the tour-based approach, the effect of year contracts on the purpose work is not included due to high correlations with the activity duration and the frequency. Month contracts have a positive effect on the utility of the purpose education. This was expected with the high share of month contracts for educational trips that was shown in the travel analysis.

The different modes have several distinct effects on the utilities of purposes. For the purpose work, trips by metro relate to a higher utility. Also multiple modes are an indication for the purpose work, although less than the metro alone. Trips made by tram have a negative relation to the utility of educational trips and a similar effect is revealed for the metro on the utility of the purpose shopping. Although interpreting the effect of mode on travel purposes is less straightforward compared to contract duration and travel frequencies, the coefficients show high similarity between the trip-based and the tour-based approach.

The departure time, or Time Of Day, was one of the main explanatory variables identified in literature. This is confirmed by the estimated coefficients, which are the highest of all categorical variables. Regarding the purpose work, the highest effect of the departure time is found in the early morning. Although not many trips are made during this time of day, the share of work trips is very large. Also the peak periods result in an increased utility, but the effect is lower, especially in the evening peak. This can be attributed to the fact that the evening peak of commuters is more spread over time compared to the morning peak. Regarding the purpose education, return trips are generally made before the evening peak. This results in an increased utility for the morning peak and a decreased utility for the evening peak, relative to the day. In the tour-based approach, however, only the morning peak was significantly

different from the day. Since educational trips are less frequently observed in the night compared to other trips, a negative coefficient was expected. Regarding the purpose shopping, all coefficients are negative, indicating that most shopping trips are made during the day.

An increasing number of trip legs decreases the utility of the purpose work. This is counterintuitive as work trips make relatively more transfers. The attribute was included since it improved the model fit, but it is also correlated to the alternative specific constant. For interpretational reasons, it might be better to exclude the effect of the number of trip legs for the purpose work. The negative effect on the purpose shopping does comply with the observations from the travel analysis.

The activity duration was expected to provide additional explanatory value to the distinction between purposes work and education. However, the implementation of the activity duration as continuous variable resulted in an insignificant effect on the purpose education. Since work trips generally have longer activity durations than education trips and are more frequent, a positive effect on the utility of education would overestimate the influence of long activities for education. Consequently, the distinction between work and education requires the implementation of the activity duration as a categorical variable, similar to the rule-based purpose inference. Longer activities do have a negative effect on the utility for the purpose shopping. This was expected since shopping generally does not involve long activities.

Even though the distinction between work and education trips was limited by the continuous level of the activity duration, the implementation did results in a substantially higher fit statistic for the tour-based model. Therefore, it can be concluded that the activity duration explains a different effect on the travel purpose than any other available attribute.

The non-home-based purpose inference model only incorporates three attributes. The mode and the concession did not have a significant effect on the utilities of any purpose. The remaining attributes indicate very similar effects compared to the trip-based and tour-based models. The effect of a year contract was not significant in the 2008 dataset, but was in every other year. Since it only involves a small value, and thus has a limited effect, we included in the model.

The large number of estimated parameters results in difficulties to assess the influence of specific attributes on the final probabilities of purposes. In order to get a feeling of the probability distributions over purposes, Table 25 presents four fictional trips and their respective probability distributions according the three purpose-inference models. The four trips represent a common trip for a specific purpose.

The first example trip shows a very high probability for work, indicating that the model is well able to distinguish work trips. The second example, related to the purpose education, assigns the highest share to this purpose. However, the probability is closer to the probability of the purpose work. The tour-based model does assign a higher probability for education compared to the trip-based model. Hence the inclusion of the activity duration does result in an improved inference of the education purpose, even though it is not specified in its own utility function. Example trip three indicates that the trip-based model does not assign high probabilities to the purpose shopping, as it is still lower than the probability of other. The tour-based model does

result in a higher probability. The fourth example trip indicates that the purpose other can also obtain high probabilities, although not as high as work.

Table 25: Examples of travel purpose inference

Attributes		*trip 1*	*trip 2*	*trip 3*	*trip 4*
Frequency		4	5	1	1
Concession		Amsterdam	Amsterdam	Waterland	Amsterdam
Contract duration		Year	Month	None	None
Mode		Metro	Bus	Bus	Bus
Departure time		Morning peak	Morning peak	Day	Night
Trip legs		2	1	1	1
Activity duration		540	420	90	240
Model	Probabilities	$P_{1,i}$	$P_{2,i}$	$P_{3,i}$	$P_{4,i}$
Trip-based	$P_{n,work}$	91%	41%	24%	22%
	$P_{n,education}$	4%	52%	10%	5%
	$P_{n,shopping}$	0%	0%	25%	8%
	$P_{n,other}$	4%	8%	41%	65%
Tour-based	$P_{n,work}$	89%	37%	5%	15%
	$P_{n,education}$	5%	55%	6%	9%
	$P_{n,shopping}$	0%	0%	50%	15%
	$P_{n,other}$	5%	7%	40%	61%
Non-home-based	$P_{n,work}$	67%	33%	21%	23%
	$P_{n,education}$	6%	38%	3%	4%
	$P_{n,shopping}$	1%	1%	13%	5%
	$P_{n,other}$	26%	27%	63%	69%

The non-home-based model results in substantially more spread probabilities of purposes. The non-home-based dataset does contain a substantially larger share of trips with the purpose other, compared to the datasets used for the trip-based and the tour-based models. This is reflected by the relatively high shares of the purpose other in all examples. The less sharp probability distributions indicate that the non-home-based model cannot distinguish travel purposes with the accuracy of trip-based or tour-based purpose inference models.

6.3.6 Assessment of the predictive qualities

By applying the purpose inference models onto the WROOV data, an assessment of the predictive qualities of the models per purpose has been performed. Figure 33 and Figure 34 show the probability distribution of each purpose categorized by the purposes observed in the WROOV data, respectively for the trip-based model, the tour-based model and the non-home-based trips model. These graphs indicate that the purposes *work* and *other* obtain the highest probabilities within their own category. The purposes education and shopping obtain lower probabilities within their own category. For *education* trips, the average probability of *work* is higher than the

probability of *education*. For *shopping* trips the average probability of *other* is higher than the probability of *shopping*. Hence it can be concluded that the models perform well regarding the distinction of *work* and *other* trips, but lack the ability to distinguish between *work* and *education* and between *shopping* and *other*.

The travel analysis already indicated that these categories are correlated in terms of the key variables. The addition of the contract duration and the activity duration were expected to allow for a differentiation between work and education purposes.

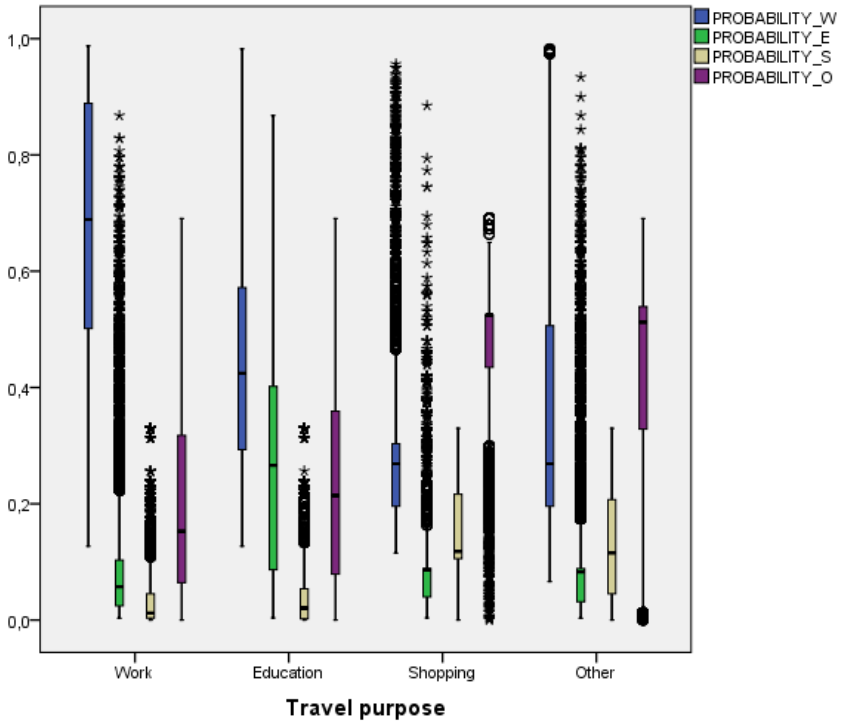


Figure 33: Probability distributions per purpose for the trip-based model

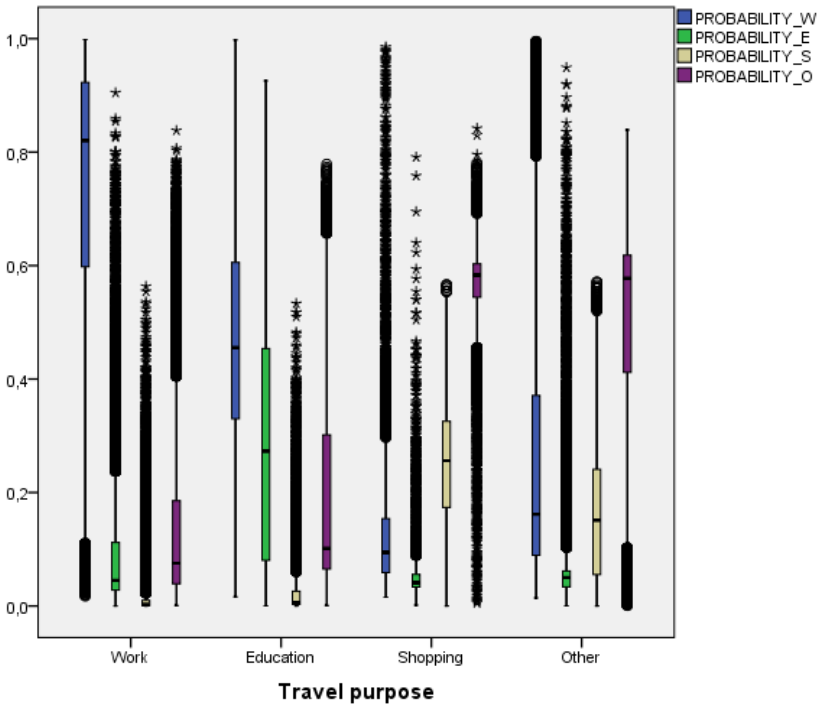


Figure 34: Probability distributions per purpose for the tour-based model

6.4 Identification of concession traversing transfers

Tests proved it was possible to add the alternative of a transfer at stops near train stations to the choice set, besides the zonal alternatives. However, this complicated the interpretation of the models, as the addition of a reference alternative added an additional value to the attributes. With the addition of the transfer alternative, the same attributes are to describe the chance a traveller originates or continues the travel in the train. Hence, the transfer to another system is positioned equal to the access or egress legs. While both processes include a trip leg, they do not describe the same phenomenon.

The method of a transfer alternative in the zonal allocation models would be interpretable with a nested logit structure. In that case, the zonal alternatives can be included in a nest, where the transfer alternative would comprise its own nest. Since this method requires a different structure of the logit models and an altered specification of the data set, this method has not been continued within this research. If future studies still have to deal with unavailability of data from one or several operators, this method might provide a potential solution.

Therefore, the identification of concession traversing transfers is based on the transfer shares found in the WROOV analysis (see Table 12 on page 59). The table presents the percentages of BTM trips using BTM as egress transport for a train leg at boarding stops near the specific train station. Similarly, the table indicates the shares of trips using BTM as access transport for a train leg at alighting stops. Using the tour-based definitions of home-end and activity-end, the percentages indicate the shares of travellers arriving in the study area by train, so using the train at the home-end, and the shares of travellers leaving the study area by train in order to perform an activity elsewhere, so using the train at the activity end of the observed BTM trip.

The tour-based definitions allow for a specification by location and, thereby, show a large difference between travellers coming to Amsterdam to perform their activities and travellers leaving Amsterdam to perform their activities elsewhere. For stations outside Amsterdam, the opposite is observed. This characterizes the attraction of the capital.

The WROOV data do not incorporate the train stations Amsterdam Science Park and Purmerend Weidevenne, since these stations did not exist at the time of data collection. As a result, transfer shares for these stations are not available. Therefore, the same transfer shares are applied as similar train stations, respectively the stations RAI and Wormerveer.

6.5 Conclusions regarding the model estimations

The estimation process of the logit models included the assessment of many model configurations in order to find the most suitable zonal allocation models and purpose inference models. This paragraph contains the conclusions of the assessment of the final model results. These conclusions on the models are also input for the final phase of this study: the matrix evaluation, which is presented in the next chapter.

6.5.1 Conclusions on the zonal allocation models

The catchment areas provide a suitable method for the alternative generation for the logit zonal allocation models. The distinction of the sizes of catchment areas by mode allows for including a larger set of relevant zonal alternatives. In general, the most relevant alternatives are selected in the choice set of the logit model as these are nearest to the used stop. However, the mode only explains part of the variation in access and egress distance distributions at stops. As a result, the selection of alternatives does not fully comply with the catchment areas of individual stops. This results in the inclusion of irrelevant alternatives, but more importantly, the exclusion of relevant alternatives. Additional fine-tuning of this method is considered possible by including attributes related to the level of service at the stop, like line frequencies and speeds, and the stop surroundings, like the stop density.

Regarding the attributes in the logit allocation models, the attribute stop density proved to be a stable estimator in the zonal allocation, in both the trip-based approach as the tour-based approach. However, this attribute is not directly related to activity locations, but can be interpreted as an indirect indicator of activity locations. It is not preferable to include indirect attributes in the model, but the attributes directly related to activity locations do not provide stable indicators in the model.

In the selections applied for the test data, filtering was based on catchment areas instead of indication of transfers. As a result respondents who entered a train station as origin or destination remained in the dataset. Since this only implies a very small share of the data, a slight overestimation of zones with train stations might occur. This overestimation is encountered in the attribute stop density, as train stations are generally served well by bus and light rail.

The tour-based approach clearly shows the difference in effect of the level of urbanization. Since addresses are the primary indicator of homes, and to lesser extent of activities, it is concluded that the tour based approach results in a more accurate allocation of home-zones. The allocation of activity zones is less accurate due to the instability of indicators of activity locations.

6.5.2 Conclusions on the purpose inference models

Five different attributes have a significant effect on the trip-based estimation of the travel purpose. The tour-based purpose inference includes two additional attributes: the activity duration and the number of trip legs. The activity duration is only included in the tour-based approach since it contains information about the relation between consecutive trips. The inclusion of the number of trip legs is due to correlations between attributes. With the inclusion of the activity duration, the effects of correlations have changed, resulting in a different model configuration.

The trip-based model performs well on the inference of the most frequently observed purposes work and other. The less frequently observed purposes education and shopping are less accurately predicted. This is caused by the correlation in key variables describing the compulsory purposes work and education, and on the other hand, the discretionary purposes shopping and other.

With the implementation of the attribute activity duration in the tour-based model, we expected a better distinction between the purposes work and education. However, the tour-based purpose inference model only shows a small accuracy increase. This can be attributed to the linear implementation. A categorical implementation of the activity duration is likely to improve the distinction between the purposes work and education.

The attributes have a different effect on distinct purposes. Therefore, purpose specific coefficients have been estimated in the models. This resulted in a large number of estimated coefficients in the model, even though the number of distinct attributes is moderate. Due to the large sample, overfitting of the model to the dataset was possible. However, the stability analysis showed that the model parameters are fairly constant over the years.

The application of the travel frequency might not be appropriate, as the distribution of travel frequency in the WROOV data does not comply with the distribution of travel frequency in the OV-chipkaart data. This is directly related to the over representation of contracts in the WROOV data that was observed in the quantitative comparison of the sources.

Land-use characteristics proved to be insignificant or instable in the logit estimation models. This can be attributed to several explanations. First, the land-use data are derived from a different time, namely 2010, where the travel data is collected between 2003 and 2009. Second, the land-use data are averaged around the used stop. The averaged characteristics do not necessarily comply with the land-use characteristics at the origin or destination, which is unknown. In their similar study, Chakirov & Erath also do not find a high explanatory value of land use characteristics.

Besides the better fit of the tour-based model, it is also preferable over the trip-based approach since it includes a higher level of behavioural richness. The tour-based approach has several qualitative advantages over the trip-based approach:

- Consistency of train transfer selection within tours;
- Application of different transfer shares at home-end and activity-end;
- Consistency between zonal allocation between consecutive trips in a tour (the destination of a trip matches the origin of the next trip);
- Addition of the highly relevant attribute of activity duration.

7 Evaluation of OD matrices

This chapter covers the evaluation of the resulting base matrices. The trip-based and tour-based models, described in the previous chapter, have been applied to the datasets of both WROOV and OV-chipkaart. In addition to these two logit modelling approaches, rule-based models have been applied. These models are based on a rule-based processing approach found in literature and adapted to fit the available data. In this chapter, we assess the qualities of these three model approaches by comparison of their resulting OD matrices.

The evaluation framework, which is described in paragraph 3.9, consists of linear regressions onto the matrix cell values. The specifications of the matrices allow for assessment of the matrices at different levels. The evaluation of the matrices at different zonal resolutions allows for the assessment of the zonal allocation models, where the distinction of the matrices by purpose allows for the assessment of the purpose inference models.

First, the procedure of the model applications is described (paragraph 7.1), in order to provide insight in the sophistication of each model. Second, the models are validated onto the WROOV data by comparison with the observed OD matrix (paragraph 7.2) in order to assess their predictive qualities. Third, in addition to the application onto WROOV data, we evaluate the differences between the model approaches by comparing the OD matrices constructed by application onto OV-chipkaart data (paragraph 0), in order to relate the quality to the required efforts of applying more sophisticated models. Fourth, the travel demand described by the different sources is compared (paragraph 7.4) in order to assess the durability of the estimated models. Finally, the conclusions of the evaluation are summarized (paragraph 7.5).

7.1 Procedure of model applications

In order to facilitate an unbiased comparison, adjustments have been made to the original data sets from WROOV and the OV-chipkaart. These adjustments are elaborated upon, before the model application procedure is expounded.

7.1.1 Creation of comparative data sets

In the qualitative comparison of the data sources WROOV and OV-chipkaart, dissimilarities have been observed in the coverage of the described system (see paragraph 3.4). In order to correct for these differences, selections have been applied to the OV-chipkaart data. Students are filtered based on their specific card type. Conversely, the filtering of tourists proved more difficult. By filtering the short term contracts, the datasets do become more similar regarding their coverage of travel products, but still do not describe the same travel in terms of key variables. The quantitative comparison shows that the OV-chipkaart data described more trips during off peak hours and more trips with shorter activities (seep paragraph 5.4).

Furthermore, the unavailability of data from adjacent public transport concessions results in the incapability to derive travel demand to zones outside the available service area. Therefore, transfers to the train network have been inferred based on transfer shares at individual train stations. In the WROOV data set, only trips without the indication of a train transfer are selected.

The total number of trips within the matrix differs between model approaches onto OV-chipkaart data due to different filtering of transfers to the train. Concession traversing transfers are inferred based on the transfer shares at train stations found in the WROOV data (see Table 12). In the trip-based approach, these shares are filtered randomly over trips boarding, respectively alighting, at the specific train stations. In the tour-based approach, filtering of concession traversing transfers at train stations is applied consistent between consecutive trips. This means that when a concession transfer is inferred at the activity-end of the trip, both the away trip and the return trip to that activity are filtered.

Table 26: Multiplication factors of OD cells for equal trip totals

Data source	*construction method*	*Trip total*	*Mean cell value*	*Standard deviation*	*Multiplication factor*
OV-chipkaart	Reference model	1.491.993	8,34	63,98	0,2148
	Trip-based model	1.587.888	8,87	39,99	0,2019
	Tour-based model	1.602.765	8,95	50,35	0,2000
WROOV	Observed	153.923	0,86	3,93	2,0826
	Reference model	156.611	0,87	5,15	2,0468
	Trip-based model	156.654	0,88	3,34	2,0462
	Tour-based model	158.718	0,89	2,93	2,0196

Furthermore, the reference models of both datasets result in different trip totals compared to the matrices constructed by the model applications due to the filtering of trips with origin or destinations with a mismatch onto the zonal grid. Some stops are projected on the wrong side of the zonal border, and in case there is no adjacent zone, i.e. water, it is filtered from the matrix. In order to compensate for the different trip totals, all OD matrices for the complete day and all purposes are corrected with a multiplication factor. The multiplication factors for OV-chipkaart matrices also include the aggregation of a week data into an average working day. Since the tour-based construction of OD matrices with OV-chipkaart data is considered to be the best representation of the real travel demand, all matrices are factored to that total. Matrices with distinctions per time of day or purpose are multiplied with the same factor as the corresponding complete matrix in order to preserve the differences between the model approaches and data sources. The multiplication factors are presented in Table 26.

7.1.2 Model application

The three applied approaches for the construction of purpose-specific OD matrices differ in complexity and, therefore, in the effort required to apply them.

The reference models require the least effort. The zonal allocation is a straight match between the used stops and zones. Because the coordinates of public transport stops and the zonal borders are derived from different systems, with a different coordinate system, slight errors are possible. However, these errors are generally negligible. Only in case the stop coordinates are on the waterfront, this can result in a mismatch. The purpose is inferred based on the activity duration and allocated to all trips in the tour.

The trip-based models require significantly more effort in their application. The allocation of origin and destination zones and the inference of the travel purpose are

based on the probabilities estimated by the logit models. This approach was chosen over a simulation procedure, which might result in biased allocations due to correlations between error terms. The allocation based on probabilities resulted in the spreading of one trip over multiple zones at the origin, multiple zones at the destination and four travel purposes. This procedure increases the data file size with a factor 35 and requires long computation times, especially with a large data set as the OV-chipkaart data.

The application of tour-based models requires the most effort, due to the consistency between trips in the same tour. This is especially complicated for tours with more than two trips. First, the selection of train transfers requires consistency. In case a transfer is inferred at the home or activity location, both the away trip and the return trip have been filtered. Second, zonal allocations for consecutive trips are consistent with the previous zonal allocation, with exception of the last trip in the tour, which destination is allocated to the same as the origin of the first trip: the home-zone. Travel purposes have not been inferred consistently within tours, so a tour can consist of trips with different purposes. It is possible to infer the travel purpose consistently within tours, similar to the zonal allocation. The purpose of the return trip can be handled in several ways. It can be allocated to the main travel purpose, for example based on the longest activity duration, or allocated to the specific purpose of being at home. The home purpose is not distinguished in this study, but the main travel purpose is also not identified. Therefore, purposes are inferred for individual trips.

The three approaches have been applied to both the WROOV data and the OV-chipkaart data. Together with the OD matrix observed by WROOV, this resulted in the seven differently constructed matrices presented in Table 26.

For the matrix evaluation, the entire matrix of the VENOM study area has been applied, although it contains many empty cells due to the structural boundaries of the available concessions. This is also visualized by the stop locations of the available concessions Amsterdam and Waterland in Figure 27.

7.2 Model validation on WROOV data

A ground truth of the real travel demand at the time of collection of the OV-chipkaart data is not available. The OD matrix observed by WROOV data can be considered as a valid representation of the travel demand. Although its accuracy is limited by the sample size and the data collection period is spread over seven years, it is the best representation of travel demand available. Therefore, the models have been validated by application onto WROOV data and comparing the resulting OD matrices with the OD matrix observed by WROOV. Five distinct specification of OD matrices have been compared for all three construction methods, on three levels of zonal resolution. Consequently, 45 comparisons have been made (see Appendix D). The resulting matrices of three approaches are assessed successively, in order of increasing complexity.

7.2.1 Rule-based OD matrices

The assessment of the rule-based approach on r^2 statistics of the linear regression shows a large deviation between the zonal resolutions. On the level of VENOM zones, which is the smallest grid in the assessment and the applied level of resolution in the VENOM model, the fit of the total OD matrix constructed with the rule-based approach is poor (r^2 = 0.330). With increasing zonal sizes, the fit increases. At the level of PC3 zones, the r^2 statistic of the total matrix approaches 1 (r^2 = 0.981), indicating a very

good fit. Consequently, we conclude that the access and egress legs have to be taken into account when describing the public transport travel demand at the resolution of VENOM zones or PC4 zones. When describing the travel demand at the resolution of PC3 zones, on the other hand, the influence of access and egress legs is insignificant regarding the allocation of origin and destination zones.

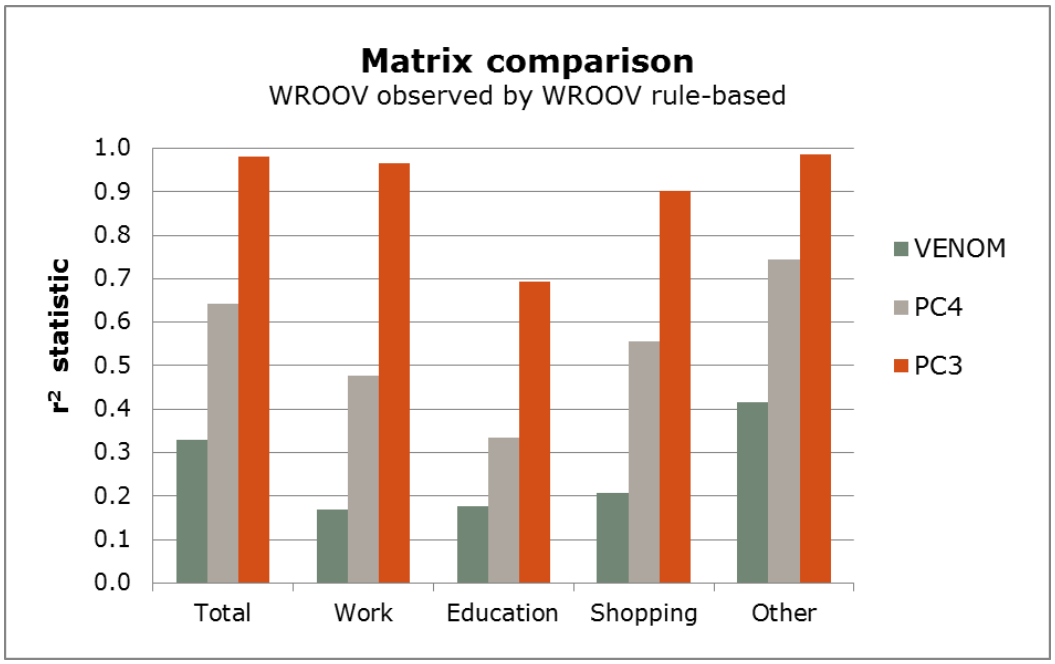


Figure 35: r^2 statistics of the rule-based model validation

The fit of purpose-specific matrices show a similar pattern as the total matrix, with low r^2 statistics at the VENOM zone level and increasing r^2 statistics for larger zones. However, the three specifically identified purposes *work* (r^2 = 0.169), *education* (r^2 = 0.177) and *shopping* (r^2 = 0.206) have a poorer fit than the non-specific purpose *other* (r^2 = 0.415). This indicates that the rule-based inference of the travel purpose does not capture the main features of the public transport travel demand.

7.2.2 Trip-based OD matrices

The trip-based approach can be considered as a more sophisticated method for the construction of purpose-specific OD matrices compared to the rule-based approach. The conversion from a stop-based matrix to OD matrix takes into account the access and egress legs by means of mode-specific catchment areas of public transport stops and allocates trip-ends to nearby origin and destination zones based on land-use attributes. The purpose inference is based on five specific trip characteristics, compared to one attribute in the rule-based approach.

The r^2 statistic of the total matrix constructed with the trip-based models shows a better fit (r^2 = 0.626) with the observed OD matrix compared to the rule-based approach at the VENOM zonal resolution. Also on the PC4 level, the trip-based models perform better than the rule-based models. On the PC3 level, the difference in r^2 statistics is very small (r^2 = 0.990).

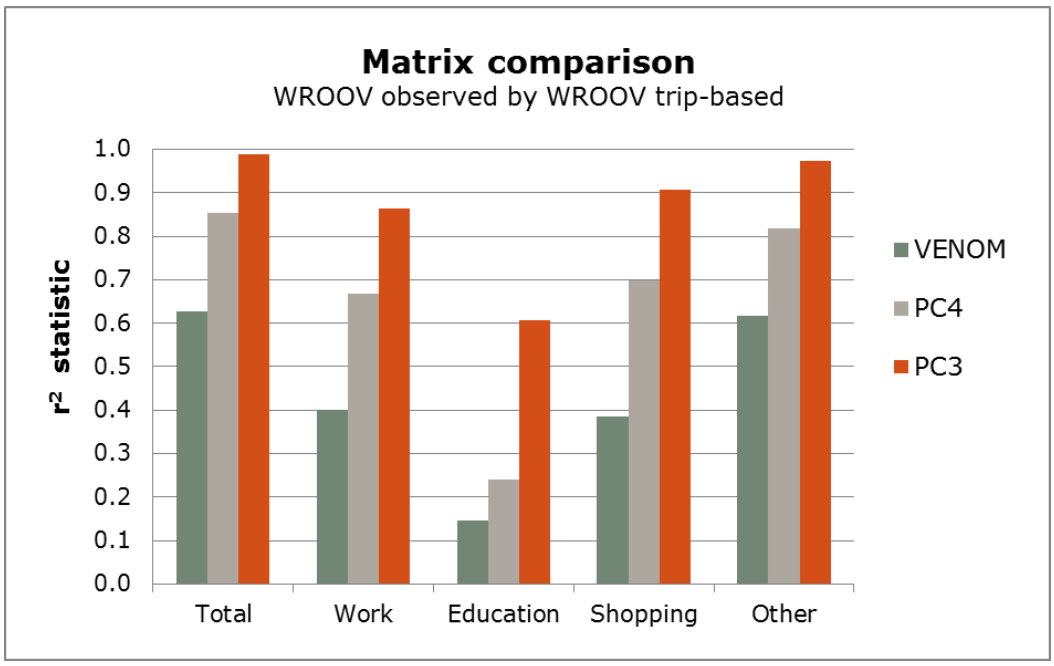


Figure 36: r^2 statistics of the trip-based model validation

Evaluating the purpose-specific OD matrices constructed with the trip-based approach, the r^2 statistics again show an inferior fit for the specific purposes *work* (r^2 = 0.402), *education* (r^2 = 0.147) and *shopping* (r^2 = 0.384), compared to the purpose *other* (r^2 = 0.618). The trip-based models perform better compared to the rule-based models on all purposes, except the purpose education, which has a slightly poorer fit. This was expected, since the model assessment indicated low probabilities for the purpose education (see paragraph 6.3.6). Conversely, the *shopping* matrix does improve compared to the rule-based approach, at a similar level as the *work* matrix, while the model assessment indicated low probabilities similar to the purpose *education*. Overall, we conclude that the trip-based approach performs substantially better than the rule-based approach.

7.2.3 Tour-based OD matrices

The tour-based models are based on the same allocation technique as the trip-based models, logit allocation, but take into account the relation between consecutive trips within tours. Specifically, the tour-based models specify trip-ends by their geographical location and allocate consecutive trips consistently. The tour-based purpose inference model incorporates the attribute activity duration, which is not available in the trip-based approach.

The r^2 statistic of the total matrix at the VENOM level is comparable to the trip-based approach, but slightly lower (r^2 = 0.574). This also holds for the total matrix at the lower levels of zonal resolution. This leads to conclude that the zonal allocation models do not perform better than the trip-based zonal allocation. This can be attributed to the lacking of land-use attributes in the tour-based zonal allocation models, which could possibly have translated the additional information of specific geographical locations into better allocation probabilities. However, these attributes are lacking since they have not been found stable indicators of home and activity zones (see paragraph 6.2.6).

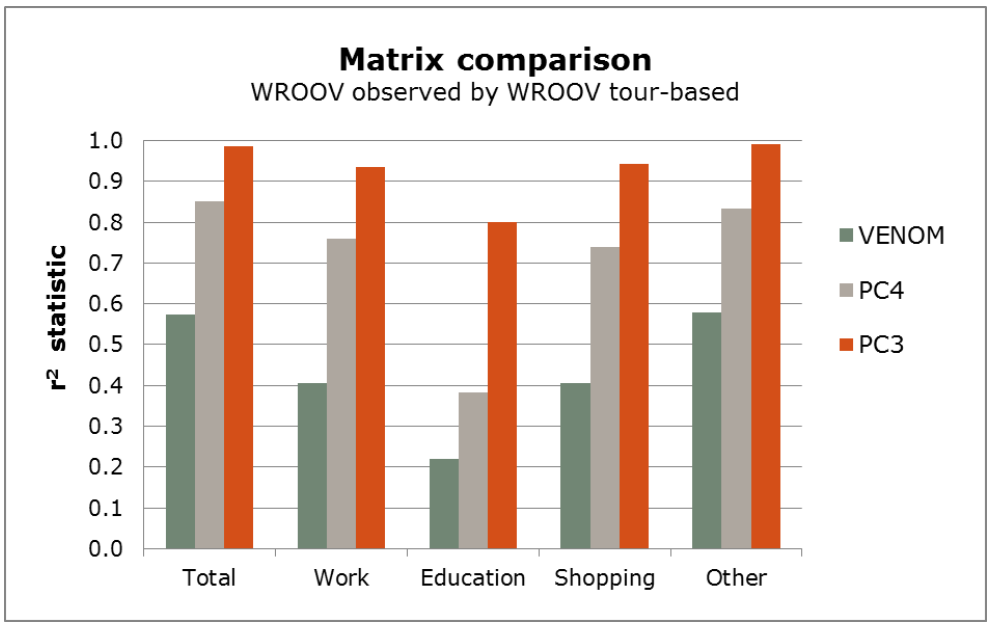
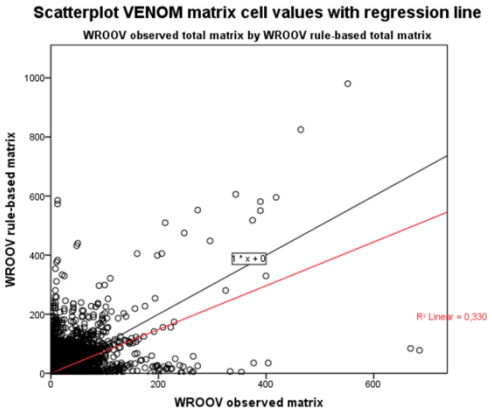
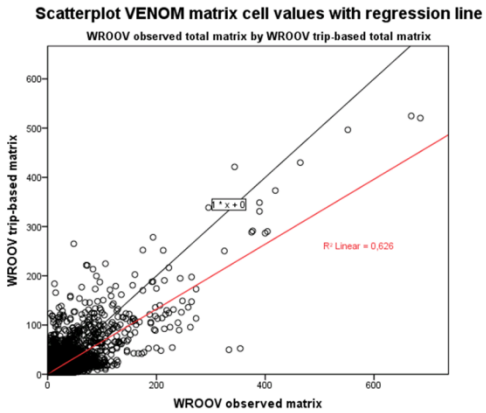


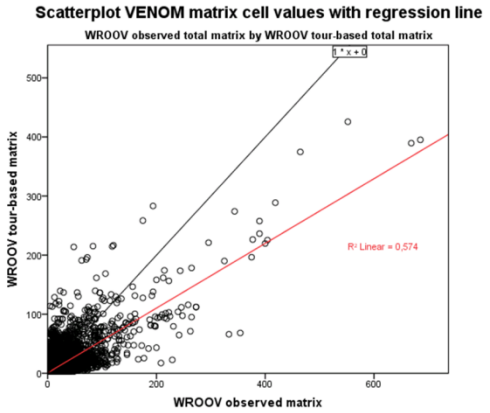
Figure 37: r^2 statistics of the tour-based model validation

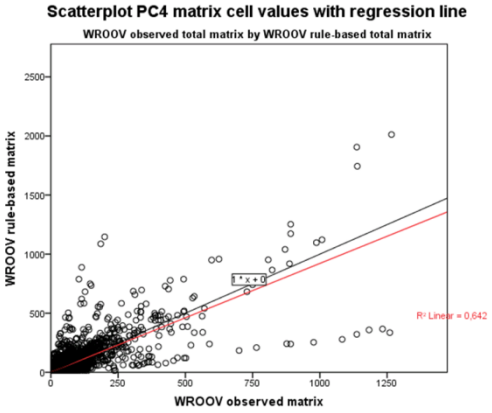
Evaluating the purpose-specific OD matrices created with the tour-based models, we find similar fit statistics for the *work* matrix at the VENOM resolution (r^2 = 0.405), compared to the trip-based models. The matrices for the purposes *education* (r^2 = 0.220) and *shopping* (r^2 = 0.404) have a better fit than their trip-based counterparts, while the *other* matrix has a slightly lower fit statistic (r^2 = 0.579). At the PC4 level, the differences between the tour-based and the trip-based matrices are more apparent, in favour of the tour-based matrices. This can be attributed to the slightly lower accuracy of the zonal allocation in the tour-based approach, which is moderated at a lower zonal resolution. The tour-based other matrix also has a better fit than the trip-based other matrix at the PC4 level. Therefore, it can be concluded that the tour-based models perform best on the purpose inference. Nonetheless, the zonal allocation does not improve by the geographical distinction of trip-ends as might be expected.

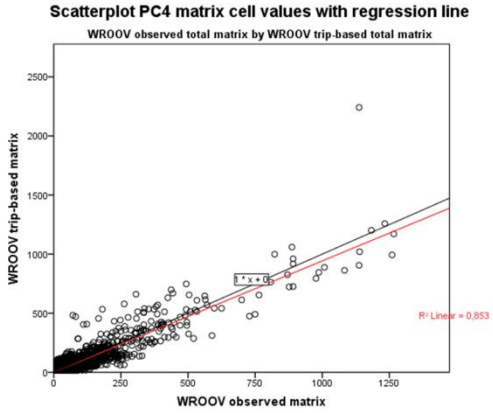
Figure 38: Model validation regression lines

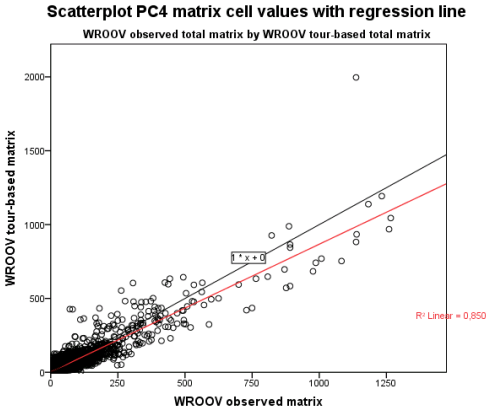

Scatterplot VENOM matrix cell values with regression line
WROOV observed total matrix by WROOV rule-based total matrix
WROOV rule-based matrix
WROOV observed matrix
1 * x + 0
R² Linear = 0,330

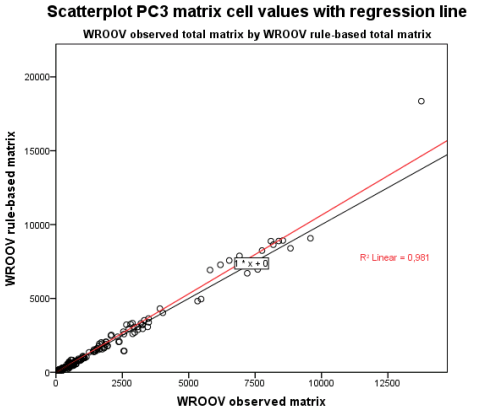

Scatterplot VENOM matrix cell values with regression line
WROOV observed total matrix by WROOV trip-based total matrix
WROOV trip-based matrix
WROOV observed matrix
1 * x + 0
R² Linear = 0,626

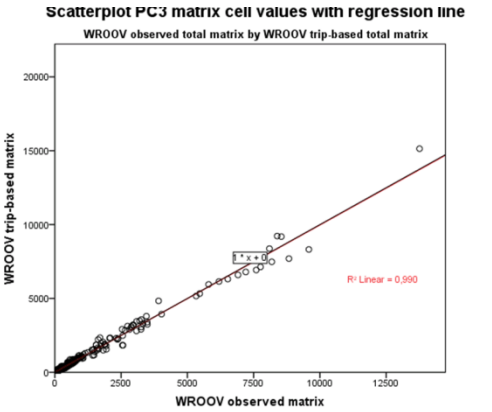

Scatterplot VENOM matrix cell values with regression line
WROOV observed total matrix by WROOV tour-based total matrix
WROOV tour-based matrix
WROOV observed matrix
1 * x + 0
R² Linear = 0,574

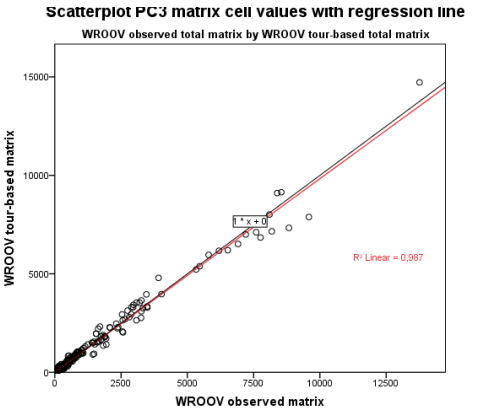

Scatterplot PC4 matrix cell values with regression line
WROOV observed total matrix by WROOV rule-based total matrix
WROOV rule-based matrix
WROOV observed matrix
1 * x + 0
R² Linear = 0,642


Scatterplot PC4 matrix cell values with regression line
WROOV observed total matrix by WROOV trip-based total matrix
WROOV trip-based matrix
WROOV observed matrix
1 * x + 0
R² Linear = 0,853


Scatterplot PC4 matrix cell values with regression line
WROOV observed total matrix by WROOV tour-based total matrix
WROOV tour-based matrix
WROOV observed matrix
1 * x + 0
R² Linear = 0,850


Scatterplot PC3 matrix cell values with regression line
WROOV observed total matrix by WROOV rule-based total matrix
WROOV rule-based matrix
WROOV observed matrix
1 * x + 0
R² Linear = 0,981


Scatterplot PC3 matrix cell values with regression line
WROOV observed total matrix by WROOV trip-based total matrix
WROOV trip-based matrix
WROOV observed matrix
1 * x + 0
R² Linear = 0,990


Scatterplot PC3 matrix cell values with regression line
WROOV observed total matrix by WROOV tour-based total matrix
WROOV tour-based matrix
WROOV observed matrix
1 * x + 0
R² Linear = 0,987

7.3 Evaluation of differences between modelling approaches

In addition to the model validation onto WROOV data, the three model approaches have been applied to OV-chipkaart data in order to evaluate the differences between the OV-chipkaart OD matrices per model approach. Since there is no observed OD matrix available for the OV-chipkaart data, like there is for the WROOV data, we assess the differences between the matrices constructed by different approaches. In this comparison, the tour-based matrices have been used as dependent and are respectively estimated by rule-based and the trip-based matrices. The matrix comparisons for the model approach evaluation consist of five distinct matrices that have been compared on three levels of zonal resolution, resulting in 30 matrix comparisons (see Appendix D).

Evaluating the differences between the total matrices of the three approaches, we find large differences in the fit statistics. At the VENOM zonal resolution, the rule-based matrix has a very low fit with the tour-based matrix (r^2 = 0.120). The trip-based matrix has a better fit, but still deviates significantly from the tour-based matrix (r^2 = 0.523). At the level of PC3 zones, the difference between the rule-based and the trip-based approach is less apparent. The fact that even at the PC3 level, the matrices do not have r^2 statistics that approach the value of 1 indicates a large difference due to the filtering of concession traversing transfers.

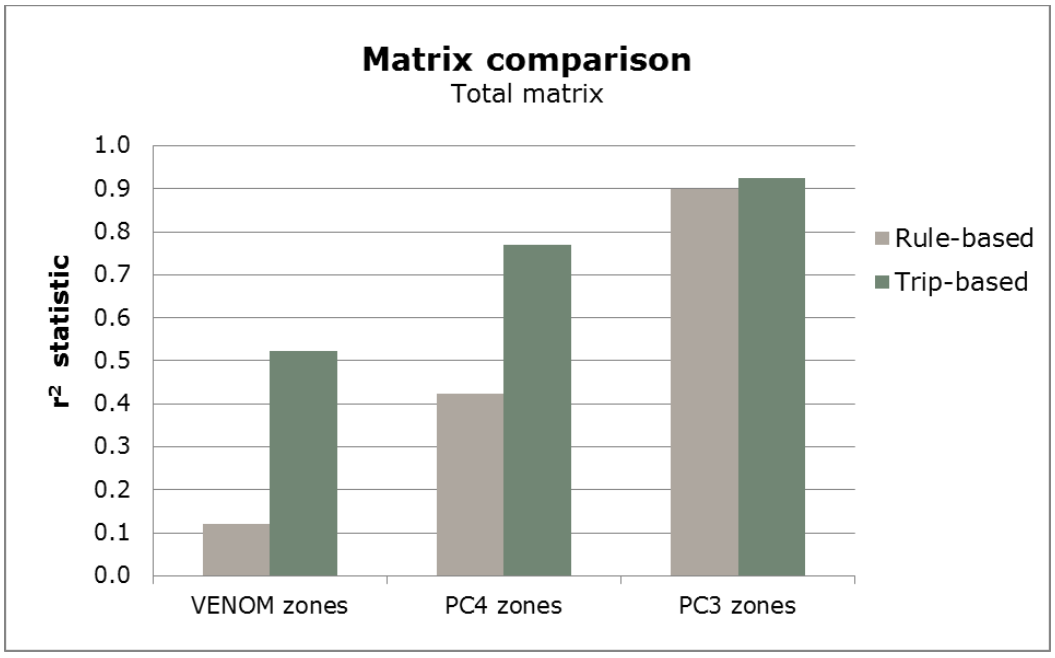


Figure 39: r^2 statistics of approach comparison of OV-chipkaart total matrices

Since a ground truth is lacking for the OV-chipkaart matrices, we have analysed the largest relations in the study area. A selection of the most frequent OD pairs, the highest values in the matrix, shows large dissimilarities between the model approaches. Even between the trip-based and the tour-based approach the selection of the highest values only show moderate comparability. Zones with train stations are very well represented in this selection of OD pairs. Hence, the influence of the selections at train stations is substantial.

The fact that, between the trip-based and the tour-based approach, the comparability of the purposes work and other is lower than the comparability of the purposes education and shopping was not expected. This might be related to the homoscedasticity assumption of linear regression. This assumption is not met by the matrix comparisons, as these show larger deviation for higher values (see Figure 38).

Especially at the resolution of VENOM zones, these deviations indicate that the current zonal allocation models do not fully capture the essence of access and egress behaviour.

7.4 Source comparison on travel patterns

The comparisons of the matrices from different sources consist of eight distinct matrices that have been compared on three levels of zonal resolution, resulting in 24 matrix comparisons (see Appendix D).

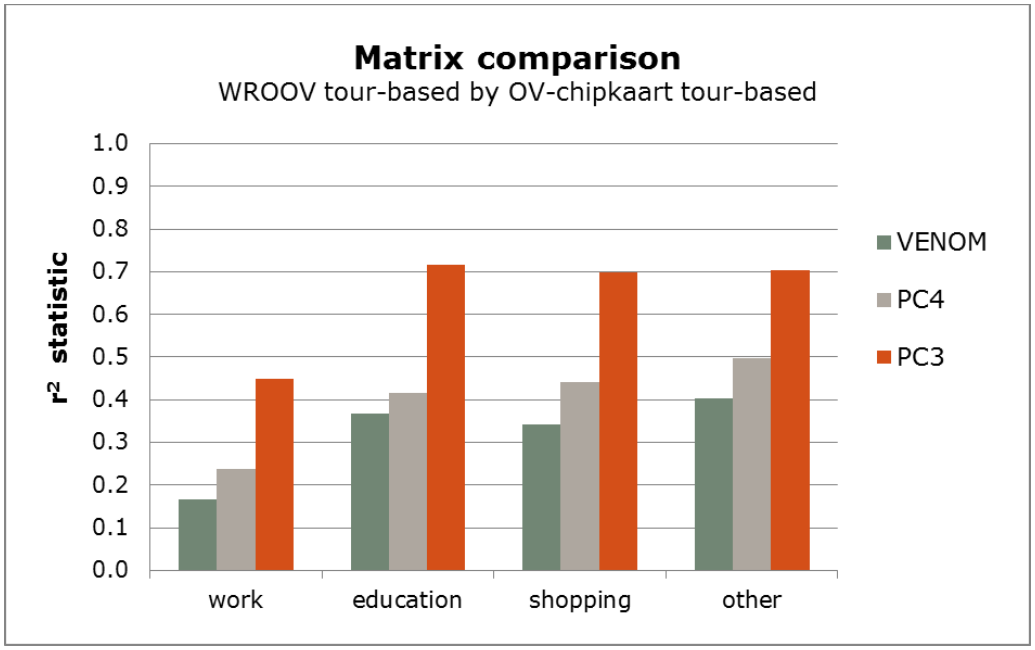


Figure 40: r^2 statistics of source comparison per purpose

The travel patterns described differs between WROOV and OV-chipkaart data. Explanations of these differences can be classified into four categories:

1. Differences in the target population;
5. The time gap between data collection;
6. Differences in the data samples;
7. Model inaccuracies.

In order to deal with differences of the first category, the differences in the target population, we have applied filters to create comparative data sets. Both students and tourist have been filtered from the OV-chipkaart data for the comparison. However, the accuracy of these filters is not perfect. The identification of students is accurate thanks to the specific card type, but the identification of tourists and concession-traversing transfers is less direct. Tourists are filtered based on short-term contracts,

but tourist may also travel with stored value. In that case, they cannot be distinguished and thus cannot be filtered. The identification of concession traversing transfers has been applied on an aggregated level per train station. The selection is based on transfers with the NS only, which leaves the transfers to the adjacent concessions of Amstelland-Meerlanden and Zaanstreek. As a result, the numbers of trips originating or terminating in the study area are overestimated.

Regarding the second category, differences due to the time gap, differences can be caused due to changes in the network and the provided transport services. Alterations of the travel supply influence the demand. In addition, the travel demand changes over time due to changes in the demography and travel behaviour. These changes have not been investigated within this research. With a minimal time gap of five years, their influence is considered small, but might not be negligible.

The third category, changes in the data sample, consists of a sample bias in WROOV and differences in the sample period. The WROOV surveys aimed at a description of travel demand for one year by means of multiplication factors. These factors have not been used in the analysis because of the uncertain influence of the many data selections. In addition, the WROOV sample consists of the stacked data of seven years. On the other hand, the OV-chipkaart data sample contains one week. Even though this week appears to be similar to the average work week, based on aggregated travel characteristics, it might be that this specific week describes different movement patterns between zones than the average work week.

The fourth category covers the model inaccuracies. As described in the previous chapter, the predictive qualities of the models vary.

7.5 Conclusions regarding the matrix evaluation

The tour-based models perform best on the model validation on WROOV data on almost all specifications of the OD matrix and all zonal aggregation levels and all assessment parameters.

The differences between tour-based models and trip-based models are small. The most notable difference occurs in the OD matrix with trips of the purpose education. For the OD matrix with trips of the purpose other, the trip-based approach even performs better on the r^2 statistic.

The rule-based approach generally performs significantly less on all assessment parameters compared to the trip-based and tour-based modelling approaches.

The influence of the zonal resolution heavily influences the r^2 statistic. The aggregation to larger zones diminishes the differences at higher resolution. However, the parameters a and b do not improve for zones in a lower resolution (i.e. larger zones). This is related to the higher mean of the cell values.

The education matrices overall have an inferior fit compared to other purposes. On the one hand, this was expected since the inference of the purpose *education* proved to be difficult with the available data. The WROOV data does not include college students, but only incorporates students of primary, secondary and vocational schools. Moreover, the locations of student places did not prove to be stable indicators of the purpose education. On the other hand, the predictive qualities are similar to the purpose *shopping*, while these matrices show higher comparability with the observed

WROOV matrix. This leads us to conclude that *educational* trips have very different spatial patterns than *work* trips, which are not captured by the trip-based model. *Shopping* trips, on the other hand, seem to have similar patterns to the purpose *other*.

8 Conclusions

The previous chapters all have contributed to answering the sub-questions of this research regarding their distinct topics. This chapter summarizes the main conclusions drawn from the different phases of this study, leading to the answer of the main research question.

The main goal of this study was the construction of purpose-specific OD matrices using public transport smart card data. We have investigated to what extent this is possible by enriching trips observed by smart card data with the required information: the origin zone, the destination zone and the travel purpose, based on data from the WROOV survey. The research was structured by dividing this main research question into four sub-questions, relating to the following subjects:

1. The identification of attributes correlated to the lacking information;
2. The transferability of information between the available sources by means of these attributes;
3. The quality of the enrichment models;
4. The evaluation of the constructed OD matrices.

First, we answer these four sub-questions, before we reflect on their accumulated conclusions in order to answer the main research question.

8.1 Relevant attributes

"Which travel characteristics are correlated to the information to be added to OV-chipkaart data and to what extent?"

8.1.1 Access and egress trip legs

First, we consider the allocation of origin and destination zones. In order to convert the stop-based matrices, which are derived from OV-chipkaart data, into OD matrices, the access and egress trip legs have to be considered, which are not observed by the OV-chipkaart.

Since we aimed for a generic conversion applicable for all stops, access and egress trip legs have been analysed in the WROOV data through their distances, as it was not feasible to assess the large number of stops individually. Previous studies indicated a relation between the access and egress distances to public transport stops and the level of service provided at these stops. However, neither the WROOV data nor the OV-chipkaart data contain information about frequency or operational speed of transit lines. Therefore the correlation of the mode has been investigated, as the level of service differs per mode.

The analysis verified that, on average, travellers cover longer distances during access and egress legs for trips by metro, and to lesser extent by tram, compared to trips by bus. Although the level of service deviates less strongly between the modes bus and tram compared to the level of service of the metro, travellers did cover larger distances at the home-end of trips to and from tram stops compared to bus stops. This difference was not observed at the activity-end. Considering attributes unavailable in OV-chipkaart data, we conclude that, at the home-end, the access and

egress distances depend mainly on the level of service. Conversely, on the activity-end, the access and egress distances depend less on the level of service and more on traveller characteristics. Moreover, the access and egress distances at the activity-end are correlated to the travel purpose.

8.1.2 Travel purpose

Considering the attributes related to the travel purpose, we have analysed the distribution of purposes over individual travel characteristics in the WROOV data. This analysis has shown that the travel purpose is strongly correlated to various travel characteristics, most notably to the activity duration, the departure time, the travel distance, the contract duration and the travel frequency. In addition, the travel purpose is also strongly correlated to the traveller characteristics unavailable in OV-chipkaart data, especially age, and to lesser extent gender.

The deviation of purpose shares over the departure time and the travel frequency distinguishes two clusters of purposes: the compulsory purposes work and education and the discretionary purposes shopping and other. The activity duration and the contract duration do show distinctions between the purposes work and education and to lesser extent between the purposes shopping and other.

Overall, we conclude that the analysed travel characteristics have large potential to describe the travel purpose. Especially the distinction between the compulsory purposes work and education and the discretionary purposes shopping and other is demonstrated by many attributes. In addition, the distinction between work and education is demonstrated by the attributes activity duration and contract duration. The distinction between the purposes shopping and other is less apparent via the investigated attributes.

8.2 Transferability of information

"How do the data sources OV-chipkaart and WROOV compare to each other?"

8.2.1 Qualitative comparison of the data sources

The qualitative comparison of the sources shows several influential dissimilarities. Predominantly, the data collection period of the WROOV surveys covers seven years between 2003 and 2009, where the OV-chipkaart data used in the analysis covers one week in 2014. This implies differences between the travel demand, which might have changed between 2009 and 2014. Moreover, the selected week of OV-chipkaart data might not be representable for the average working week. Even though the selection of the week was based on aggregated similarity with the full year average, circumstances, like weather conditions or maintenance, might influence the depicted travel patterns.

The passive collection of OV-chipkaart data results in lacking information for the construction of OD matrices. On the other hand, the passive data collection also facilitates a sample that approximates complete coverage of the entire public transport system. The only exceptions are fare dodgers and forgotten check-ins and check-outs. Despite this high potential coverage, only data from two public transport concessions were available for this study, resulting in unobserved transfers to adjacent concessions. Since these transfers result in different origins or destinations, this is a limitation to this study due to the unavailability of data. The consequences of this limitation on travel are discussed under sub-question 4 (paragraph 8.4).

On the other hand, the WROOV data do not cover the complete public transport system, but lack student cards and local tickets, which were mainly used by international tourists. Moreover, the WROOV surveys focus on bus and light rail and do not include train travel, although transfers to the train are indicated. Therefore, the WROOV data is not fit to estimate the lacking information for trips made by students and international tourists. This discrepancy in the coverage of both sources has been corrected by removing these travellers from the OV-chipkaart data. Students were easily identifiable by their card type but international tourists are identified indirectly, based on short term contracts, day and week tickets, since they are more likely to stay in Amsterdam for a short period.

8.2.2 Key variables

By comparing the available information in both sources, key attributes have been identified that are available in both sources. Since the WROOV survey is focussed on public transport, the data includes many attributes that are also available in OV-chipkaart data, resulting in a large set of potential attributes that can be used to describe the lacking information in OV-chipkaart data. Table 27 presents the key variables and their representation in the purpose inference models. The next paragraph discusses their appropriateness as model attributes

Table 27: Key variables and their representation in the purpose inference models

Key variables between OV-chipkaart and WROOV data	*Representation in the purpose inference models*	*Unit/Categories*
Activity duration	Continuous	Minutes
Travel frequency	Continuous	Travels per week
Departure time	Categorical (5)	Early morning Morning peak Midday Evening peak Night
Travel distance	Not included	(kilometres)
Contract duration	Categorical (3)	Year Month None
Fare	Not included	(full fare Reduced fare Unlimited travel)
Mode	Categorical (4)	Bus Tram Metro multiple
Concession	Categorical (3)	Amsterdam Waterland Both
number of legs within trip	Continuous	Legs per trip
number of trips within tour	Not included	(Trips per tour)

8.2.3 Quantitative comparison

The quantitative comparison on these key-variables shows slight deviations between WROOV and the OV-chipkaart, i.e. after filtering of students (20%) and international tourists (9%) from the OV-chipkaart data. Regarding the modal shares, related to the access and egress distances, the WROOV data slightly overestimates the shares of multiple modes in one trip at the expense of the modes tram and metro. Regarding the attributes related to the travel purpose, the WROOV data underestimate trips with activities shorter than three hours and overestimate activities longer than nine hours. Moreover, the WROOV data underestimate short distance trips. These discrepancies verify this well-known phenomenon of underreporting of short trips in travel surveys.
In addition to the overestimation of long activities, the WROOV data also overestimate the trip shares in peak hours at the expense of trips during off-peak hours and the higher travel frequencies. These three discrepancies are correlated to the higher share of long-term contracts in the WROOV data, compared to the OV-chipkaart data.

On the other hand, the OV-chipkaart data underestimate short activities. The distinction between short activities and transfers relies on rule-based processing. This study applied an enhanced framework of processing rules for the identification of short activities. Compared to the applied processing rules, the standard application of 35 minutes transfer time overestimates transfers with 22%. Nonetheless, the distribution of activity durations still indicates an underestimation of short activities.

This leads to the conclusion that the WROOV data over represents long term contracts compared to the current travel demand, and consequently, the purposes work and education. This can be caused by the data collection method or changes in the travel demand. Because the exact cause cannot be determined we have applied the described attributes in the model estimation despite their dissimilarities.

The key variables fare and number of trips within a tour have not been applied in the enrichment model, as they are not considered stable indicators of the travel purpose. Fare systems have been altered since the introduction of the OV-chipkaart and the WROOV data does not contain tours with more than two trips, while in the OV-chipkaart data 23% of the tours have more than two trips.

8.3 Quality of the enrichment models

"How can OD matrices by purpose and by time of day for the mode BTM be constructed using OV-chipkaart data?"

Three model approaches have been applied for the construction of purpose-specific OD matrices with OV-chipkaart data. The rule-based approach does not require additional data and can be applied based on simple assumptions. The trip-based and the tour-based models do require survey data in order to estimate the model parameters that represent the influence of the selected key variables. The tour-based models require more information than the trip-based models, as they incorporate the interaction between consecutive trips within tours. The trip-based approach describes travel between activities at origins and destinations, and thus defines trip-ends by their start and end points. The tour-based approach describes travel from the home-location back home. Hence, tours consist of one or multiple consecutive trips, depending on the number of activities performed. Consequently, trip-ends are classified by their location, either the home-end or activity-end, instead of origins and destinations.

8.3.1 Zonal allocation models

The rule-based zonal allocation allocates origins and destination based on stop locations. This method requires little effort and no additional data, but results in inaccurate allocations to zones in high resolution zonal grids.

The trip-based zonal allocation involves a logit allocation procedure. The available alternatives are selected based on the mode-specific catchment areas. Trips are allocated based on the probabilities of the available zonal alternatives. Separate models were estimated for the origin allocation and the destination allocation, but since 90% of the data consists of tours with two trips, the models are nearly equal. Three attributes proved to have a significant explanatory value to the origin and destination zones:

- The share of the catchment area in the zone;
- The stop density and;
- The level of urbanization.

The latter has a relatively low t-value and the stability analysis over the separate WROOV years showed relatively large deviations in the influence of the urbanization level. However, the level of urbanization was significant in every year and therefore included in the allocation models for both origin and destination.

The tour-based zonal allocation models deviate from the trip-based models in their definition of trip-ends. Due to the geographical distinction of home-ends and activity-ends, the influence of land-use attributes was expected to be greater compared to the trip-based model. However, this hypothesis was not verified, as land-use attributes number of jobs and student places were found to be unstable indicators of the activity zone. The distinction between home and activities instead of origins and destination did indicate the cause of the deviation in influence of the urbanization level in the trip-based models. At the home-end, the influence of urbanization level on the utility of zones tripled, while on the activity-end it was not found to be a stable indicator.

Several causes can be identified for the unforeseen lack of effect of land-use attributes. Primarily, the zonal data relate to the year 2010, the base year of the current VENOM model, while the travel data is collected between 2003 and 2009. Another possible explanation is that the land-use data do not differ enough between the available zones. Since the zones are relatively small, adjacent zones are likely to have similar aggregated land-use characteristics.

The stop-density does prove to be a stable indicator of home and activity zones, although this attribute is only indirectly related to trip production and attraction. Stop locations, and thus the stop density, are adapted to fit the travel demand, but also influence the demand. This interdependency of travel demand and stop density might result in overfitting of the model. Nonetheless, we included the stop density in the models as indicator of nearby activity locations, since its effect is not described by any of the other model attributes. As a result, the fit of the zonal allocation models is optimized, but, at the expense of the model durability.

8.3.2 Purpose inference models

Regarding the purpose inference models, the rule-based approach includes crude simplifications that do result in accurate shares per purpose when accumulating all trips, but at the level of OD pairs, this is not considered a suitable method.

The trip-based purpose inference model includes five trip attributes with specific influences on different purposes: the concessions travelled in, the contract duration, the travel frequency, the used modes and the departure time. The tour-based purpose inference model is similar to the trip-based model, but adds the attribute activity duration. The activity duration has a high explanatory value of the purpose *work*, resulting in a good fit of the tour-based purpose inference model.

The tour-based purpose inference model performs well on the estimation of the travel purposes *work* and *other*. However, the estimation of purposes *shopping* and *education* proves to be more demanding, these purposes cannot be identified with high accuracy by the estimated models. The purpose *education* is in several ways similar to the purpose *work*, but education does not generate as many trips. This complicates the identification of this less-frequent compulsory travel purpose. The same problem arises with the estimation of the purpose *shopping*, which is similar to the more-frequent purpose *other*. These clusters already emerged during the identification of relevant attributes, but with the inclusion of the activity duration and the contract duration a more accurate distinction between *work* and *educational* trips was expected. However, in the current form the purpose inference model does not perform well on the identification of the purposes education and shopping. The excellent fit statistic is mainly based on the accurate inference of the most frequent purposes *work* and *other*.

The quantitative comparison of key variables in the employed data sources also indicated dissimilarities between the descriptions of both sources regarding several attributes applied in the purpose inference model. Most notably, the distributions of contract duration and travel frequencies do not match. This might result in a model that is over fitted to the survey data and, consequently, overestimates the shares of *work* trips.

The logit purpose inference models do not incorporate land-used characteristics since these did not prove to be stable and significant attributes in the purpose inference. Land-use characteristics are often mentioned as explanatory variables of the travel purpose in the literature, but the most comparable study by Chakirov & Erath (2012) also indicates very little influence of land-use characteristics in a logit allocation model. Possible explanations for the insignificance of land-use attributes are the fact that these do not originate from the same year as the travel data. Furthermore, the land-use characteristics used in the model estimations are the aggregated values around the used stops. Since the actual origins and destinations are not observed in OV-chipkaart data, the aggregated land-use attributes might not comply with the actual values at the origin or destination zone.

Overall, we conclude that the models based on travel characteristics are well fit to distinguish between compulsory and discretionary trips, but do not contain sufficient power to accurately infer the purposes *education* and *shopping*.

8.4 Matrix evaluation

"How do base matrices created by different methods compare to each other?"

The three model approaches have been applied to WROOV data in order to compare the constructed matrices per approach with the observed WROOV matrix. In addition, the models have been applied to the OV-chipkaart data in order to compare the matrices constructed with the different approaches with each other.

8.4.1 Zonal allocation

The assessment of the zonal allocation models consists of the comparison of the constructed total matrices, referring to an average working day, with the observed WROOV matrix.

The comparisons of matrices constructed with WROOV data indicate an inferior fit of the rule-based matrices compared to the trip-based and tour-based matrices, which have similar fit statistics. The trip-based matrix even has a slightly better fit to the observed WROOV matrix than the tour-based matrix, which can be attributed to the inclusion of the attribute urbanization at both trip-ends.

Assessment of the matrices constructed with OV-chipkaart data shows large deviations between the rule-based and the tour-based approach. The matrix constructed with the trip-based models shows higher comparability with the tour-based matrix, but still lower than expected, taking into account their similarity in the zonal allocation.

Evaluating the differences between matrices at lower levels of resolution, we find that the added value of the probabilistic allocation of zones reduces at the level of PC3 zones. Hence, rule based processing of smart card data can provide suitable OD matrices for models with relatively large zones, comparable to PC3 areas. However, transport models with a higher level of resolution, like the VENOM model, are better served with the more deliberate approach of logit allocation models.

The fact that the trip-based and tour-based OV-chipkaart matrices have a relatively poor fit at the PC3 level indicates that these matrices are structurally different. Since the effect of access and egress distances at this level of zonal resolution is small, it can be concluded that the applied methodology of filtering trips with concession traversing transfers has a large effect on the travel demand described by these OD matrices.

8.4.2 Purpose inference

The assessment of the purpose inference models is based on the comparison of purpose-specific matrices. Compared to the generic zonal allocation models, the purpose inference models have an additional assessment option due to their specific alternatives: the comparison of the observed purposes in WROOV data with the probabilities calculated by the logit models. This comparison indicated a slight improvement in the accuracy of the tour-based model compared to the trip-based model.

When comparing the resulting matrices of the two approaches, this improvement is not visible at the level of VENOM zones, where the fit of the tour-based model is similar or poorer than the fit of the trip-based matrices onto the observed WROOV matrices. This can be ascribed to the differences in zonal allocation, which has to be taken into account when comparing OD matrices. However, at the level of PC4 zones, the tour-based matrices do indicate an improved fit with the observed matrices, especially for the purposes work and education.

Overall, the matrix comparisons verify the high accuracy of the inference of the purposes *work* and *other* and the lower accuracy of the inference of *education* and *shopping* purposes. Between the lower scoring purposes, the shopping matrix does indicate a better fit than the education matrix. This leads to the conclusion that education trips have more specific spatial patterns that deviate from commuting patterns, while shopping trips have a similar pattern as trips made for *other* purposes.

8.4.3 Durability of the method

In order to assess the durability of the method, the matrices constructed with WROOV data and OV-chipkaart data were compared. The total matrices are compared on the stop level, constructed with the rule-based approach, to eliminate the effect of access and egress legs and ensure an unbiased comparison. For the purpose-specific matrices, the comparison is made on the tour-based matrices, since this approach is the most accurate. The assessment based on linear regression shows a relatively poor fit, indicating that the described travel demand does not match. Especially, the work matrices have low comparability, which can be explained by the over-representation of work trips in the WROOV data.

While the influence of changes in travel demand may be small over time, the differences between the sources indicate that the models estimated on WROOV data have limited durability. The method of construction is renewable, but requires periodic updating of the survey data to re-appraise the model parameters.

8.5 Answer to the main research question

"To what extent can the travel purpose, origins and destinations of public transport trips derived from smart card data be inferred based on information from survey data, in order to construct purpose specific OD matrices suitable as base matrices in transport models?"

Accumulating the answers of the sub-questions, it can be concluded that the construction of purpose-specific OD matrices based on survey data generates added value to the representation of travel demand by OD matrices. The probabilistic allocation procedure outperforms the rule-based approach in the allocation to origins and destination zones. Especially on higher levels of resolution, like the zonal grid of the VENOM model, the improvement of the allocations is substantial. At a lower level of resolution, like the PC3 level, the influence of the distances covered during access and egress legs is reduced by the larger internal distances of the zones. Nonetheless, for regional transport models like the VENOM model, the logit zonal allocation method results in a significant improvement of the description of the travel demand by base matrices compared to a direct conversion of stops to zones.

In addition to the increased accuracy of the origins and destinations, the purpose inference based on survey data allows for an accurate distinction of trips between the purposes *work* and *other*. The less appearing purposes *education* and *shopping* are closely related to, respectively, the purposes *work* and *other*. Therefore the inference of these purposes is less accurate. Combining these two clusters does provide an accurate distinction between compulsory purposes and discretionary purposes. The attributes in survey data required to perform this method of enrichment include the home and activity zones, the used stops and modes, the departure times, the activity duration, the contract duration and the travel frequency.

The augmentation of the trip-based approach into the tour-based approach resulted in a more accurate inference of the travel purpose due to the inclusion of the attribute activity duration. In addition, the tour-based trip-end distinction, by home and activity side, resulted in increased effect of the level of urbanization at the home-end. Consequently, allocation of home zones is more accurate than the trip-based allocation to origins and destinations. On the other hand, the tour-based approach did not improve the allocation to activity zones, since land-use attributes jobs and student places did not prove to be stable indicators of activity zones. Overall, the tour-based contains more behavioural richness and is, therefore, preferable over the trip-based approach.

Due to limitations of the available data, the constructed purpose-specific OD matrices cannot readily be applied as base matrices in transport models. In order to realise comparable datasets, students and international tourists have been filtered from the OV-chipkaart data, since the WROOV survey does not cover these travellers. These travellers make up nearly 30% of the total travellers and therefore have to be considered in the assessment of the total public transport travel demand.

Another prominent limitation of this study is the unavailability of the OV-chipkaart data from adjacent concessions, resulting in unobserved transfers. The influence of these transfers on the OD matrix is substantial, since the origin or destination zone is not near the observed stop and many travellers transfer to the train network. This limitation is non-existent when data from all adjacent operators is available.

The applied methodology of constructing purpose-specific OD matrices based on smart card data shows great potential. During this research, an operational method has been built, which results in a more accurate description of public transport travel demand than previously available. However, restrictions of the available data and imperfections in the application have resulted in limited applicability and durability. With additional fine-tuning of this method and increased availability of the data, the method presented in this report can be enhanced to a fully applicable approach that can lead valuable improvements of the quality of public transport demand forecasts.

9 Recommendations

Based on the conclusions presented in the previous chapter, this chapter provides the recommendations to the various parties involved in this study. Two categories of recommendations are distinguished. First, recommendations for follow-up research are provided (paragraph 9.1). Subsequently, recommendations for the utilization of the results of this study are listed (paragraph 9.2).

9.1 Follow-up research

The recommendations for follow-up research are categorized by three subjects:

- Enhancements to the enrichment models;
- Expanding the method to obtain complete coverage of the public transport system;
- Related research topics that can build on this study.

9.1.1 Enhancements to the enrichment models

Reflecting on the applied methodology, we have found several issues that leave room for improvement. The enhancements are categorized into three subjects:

- Zonal alternative selection by means of catchment areas;
- Combined allocation of activity zones and corresponding purposes;
- Evaluation of the matrices.

Zonal alternative selection by means of catchment areas

The mode-specific catchment areas capture the main influence of the level of service at stops, but the distances covered during access and egress transport still show high deviation between stops of the same mode. We believe that the selection of zone alternatives in the logit zonal allocation can be improved by incorporating additional attributes related to the level of service: the frequencies and operational speeds at stops, instead of the indirect indication of the level of service via the mode. In addition, the relative location of the used stop to the locations of other stops on the same line provides additional insight in the access and egress behaviour. Adding these attributes in stop-specific catchment areas will improve the identification of the zonal alternatives and, consequently, the quality of the zonal allocation.

Combined allocation of activity zones and corresponding purposes

Regarding the probabilistic allocation of trips to the available zones, the inclusion of the stop density in the zonal allocation models should be re-assessed. Because its relative influence on the utility of alternatives is very large compared to its influence on the model fit, the inclusion of this attribute in the zonal allocation models is questionable. It is recommended to extend the search for alternative attributes that relate more directly to the trip production and attraction of the zones.

We recommend directing this search at magnification of the differences in land-use characteristics between adjacent zones. This can be achieved by subtracting the mean from the attribute values. In addition, the zonal allocation models could be estimated specifically per concession. The zonal structure in the highly urban concession Amsterdam has a higher resolution than the zonal structure in the more rural concession Waterland. Moreover, the mean values of land-use characteristics differ between Amsterdam and Waterland.

Another search direction comprises assessing the influence of the mismatch between the collection period of the travel data (2003-2009) and the reference year of the zonal data (2010). This can be done by selecting the WROOV data from the year 2004 and estimating the zonal allocation models with land-use data from 2004, which is the previous base year of the VENOM model. However, since the changes in land-use characteristics over this period are expected to be small, we recommend focussing on the magnification of differences in land-use characteristics.

Since the access and egress distances were found to be correlated to the travel purpose, we recommend estimating the activity zone with the corresponding purpose together in one model. The attributes jobs and student places directly refer to a travel purpose, respectively work and education. In the distinct purpose-inference models, these attributes did not prove valuable indicators, partly because the land-use data used did not necessarily comply with the origin or destination zone. Moreover, the presence of jobs in a zone might reduce the chance of other activities occurring in that same zone. These two issues can be solved by joining the purpose inference with the activity zone allocation. Therefore, the combined allocation of the activity zone and the purpose will have higher chance of indicating significant and stable relations. The model structure of such an integrated allocation has been set-up and successfully tested, but the optimization is complex and time-consuming.

In addition, the implementation of the attribute activity duration as categorical variable might increase its explanatory value for the purpose inference. In the current tour-based purpose inference model, the activity duration is implemented as a continuous variable, but very long activities are strongly related to the purpose work, diminishing the explanatory value of the activity duration for the purpose education.

Concluding, it is recommended to perform a re-appraisal of the home zone allocation model parameters and estimate a combined activity zone and purpose allocation model, based on a new survey that complies with the OV-chipkaart data sample.
With the knowledge acquired during this research, which consisted of many data handling procedures and conceptual ideas, the next optimisation round will require substantially less effort.

Evaluation of base matrices
Regarding the evaluation of the purpose-specific OD matrices, it is recommended to extend the evaluation by means of linear regression. We recommend assessing the comparability with the network performance. By assigning the resulting matrices to the network with a route-choice model, a direct comparison can be made with the network performance observed by OV-chipkaart data. This evaluation provides additional insight in the influence of differences between matrices on the network performance.

9.1.2 Complete coverage of the public transport system

In order to obtain complete coverage of the public transport system, the trips that have been filtered from the OV-chipkaart data, in order to comply with the coverage of the WROOV surveys, have to be taken into account. These trips include trips made by students and international tourists. In order to take into account the entire public transport system, we recommend initiating a new survey, in connection with OV-chipkaart data. Currently, travellers can review the transactions made with their OV-

chipkaart in an online overview[13]. Moreover, it is possible to add missed check-in and check-out transactions to this overview in order to reclaim part of the entry fare. In a similar way, travellers could supplement their travel purpose, origins and destinations. This requires very little time and effort and thus results in a low respondent burden. Moreover, such an online survey reduces processing times and costs.

A new survey also enables compatibility between the survey data and the smart card data. The WROOV surveys have provided a very large sample with all the information required for the enrichment of OV-chipkaart data. However, the travel patterns described by the WROOV data do not entirely correspond with the patterns described by the OV-chipkaart. This problem can only be solved by estimating the enrichment models on a representative sample of the available OV-chipkaart data.

Furthermore, renewal of the survey data is required at some point in time due to the decreasing value of information over time. Changes in the public transport system, concerning travel behaviour, demand and supply, develop slowly, but have to be taken into account. Since the WROOV surveys have been terminated after 2009, the durability of the estimated models is limited. Therefore, we recommend reassessment of the models with up-to-date, OV-chipkaart compatible survey data that cover the complete public transport system.

Similarly to survey data, smart card data require complete coverage for the construction of OD matrices. The problem with coverage of the OV-chipkaart is related to the fragmentation of data at different operators, serving specific public transport concessions. In order to construct fully applicable public transport OD matrices, the OV-chipkaart data is required from all operators with services in the study area. For the case study, these include additional data from Connexxion of the concessions Amstelland-Meerlanden and Zaanstreek and the data from NS regarding the national railways. It is advisable for public transport authorities to regulate the availability of OV-chipkaart data for research purposes, like this study, in the concessionary conditions. This case study has shown that the data from different operators can be coupled in order to obtain insight in the travel demand in the complete public transport system, beyond the structural boundaries of concessions. A complete overview of the transport demand generates added value for both public transport authorities and operators.

9.1.3 Related research topics

Smart card data provide many other opportunities for the improvement of public transport travel demand modelling than the construction of OD matrices. This study was focussed on the construction of OD matrices as first step of the incorporation of this new data source in transport modelling.

In the applied method, the construction of the OD matrix is completely separated from the construction of synthetic OD matrices by the model. This approach can also be turned around by calibrating the synthetic model based on the observed travel demand in OV-chipkaart data.

Another topic is the calibration of travel supply modelling by route choice models. Since the OV-chipkaart data also contain the exact route travelled, route choice models can be calibrated by assignment of the constructed OD matrices onto the network. This is related to the evaluation of the matrices by means of the network

[13] Available via Mijn OV-chipkaart at www.ov-chipkaart.nl.

loads. When both the matrices and the route choice models have not been validated, it cannot be determined which of the two is related to potential dissimilarities. Therefore, we recommend first to optimize the construction of OD matrices, before calibrating route-choice models. Nonetheless, route choice model calibration is a very promising research topic that could build on this study.

9.2 Utilization of the results

When completed with the travel demand of students and international tourists, the constructed purpose-specific OD matrices can be used in travel demand studies in the VENOM study area, for example the influence of the new metro line in Amsterdam, leading to an improved description of the travel demand in the current situation. After implementation in the VENOM model, this will also lead to more accurate forecasts of travel demand in the forecast year.

Both the public transport authority SRA and the operators GVB and EBS can benefit from the increased insight in the relation of travel demand between of the respective concessions. By combining their OV-chipkaart data, all three parties have acquired insight in the previously concealed travel demand of beyond the borders of individual public transport concessions.

With the continuous flow of OV-chipkaart data, it is recommended to update the base year of the public transport OD matrices more frequently. The current renewing cycle of approximately five years can result in significant differences in travel demand, while an update based on OV-chipkaart data only requires a new application procedure of the enrichments models.

Moreover, the longitudinal character of the data collection can provide increased insight in the relation between average travel demand and peak demand. This can be done at two levels: the representation of the average working day and the diffusion of demand over the day. The analysis of the model on the average working day can be extended to a bandwidth analysis by representing quiet working days and busy working days, based on the yearly average.

The construction of purpose specific OD matrices with OV-chipkaart data allows for the specification of matrices by any desired specification of time. Trip generation and trip distribution models estimate synthetic matrices for a specific speak period or for 24 hours. Matrices constructed with OV-chipkaart data allow for selections based on the time-stamp, directly available from the data. Thereby, the construction of OD matrices with OV-chipkaart data provides increased insight in the distribution of the travel demand over the day.

The comprehension of travellers' motivations to travel provides valuable information for the operators that can be used in their policies of fare schemes and ticketing systems. However, it has to be taken into account that the estimation was based on survey data with higher shares of long term contracts, and thus higher shares of commuting travellers. In addition, it has to be noted that the logit allocation to zones still incorporates substantial deviations for individual WROOV matrix cells at the VENOM level, although it describes the travel demand significantly better than matrices constructed with the rule-based approach.

Bibliography

Agard, B., Morency, C., & Trépanier, M. (2006). Mining public transport user behaviour from smart card data. *12th IFAC Symposium on Information Control Problems in Manufacturing.* Saint-Etienne: INCOM.

Agard, B., Morency, C., & Trépanier, M. (2009). *Mining Smart Card Data from an Urban Transit Network.* Montréal: École Polytechnique de Montréal, Canada.

Allos, A., Merrall, A., Smithies, R., & Fishburn, R. (2014). New data sources and data fusion. *European Transport Conference.* Frankfurt: AET.

Alshalalfah, B., & Shalaby, A. (2007). Case Study: Relationship of Walk Access Distance to Transit with Service, Travel, and Personal Characteristics. *Journal of Urban Planning and Development, 133*(2), 114-118.

Axhausen, K., Schönfelder, S., Wolf, J., Oliveira, M., & Samaga, U. (2003). 80 weeks of GPStraces: Approaches to enriching the trip information. *Institut für Verkehrsplanung und Transportsysteme.* Zürich: ETH.

Bagchi, M., & White, P. (2005). The potential of public transport smart card data. *Transport Policy, 12*(5), 464-474.

Barry, J. J., Freimer, R., & Slavin, H. (2009). Use of entry-only automatic fare collection data to estimate linked transit trips in New York City. *Transportation Research Record: Journal of the Transportation Research Board, 2112*(1), 53-61.

Ben-Akiva, M. E., & Lerman, S. R. (1985). *Discrete choice analysis: theory and application to travel demand.* Cambridge: MIT press.

Bierlaire, M. (2003). BIOGEME: a free package for the estimation of discrete choice models. *3rd STRC Swiss Transport Research Conference.* Ascona: STRC.

Blythe, P. T. (2004). Improving public transport ticketing through smart cards. *Proceedings of the ICE - Municipal Engineer, 157*(1), 47-54.

Blythe, P. T., & Holland, R. (1998). Integrated ticketing - Smart cards in transport. *IEE Colloquium: Using ITS in Public Transport and in Emergency Services.* London: IEE.

Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies, 17*(3), 285-297.

Chakirov, A., & Erath, A. (2012). Activity Identification and primary location modelling based on Smart Card payment data for Public Transport. *13th International Conference on Travel Behaviour Research .* Toronto.

Chen, C., Gong, H., Lawson, C., & Bialostozky, E. (2010). Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. *Transportation Research Part A: Policy and Practice, 44*(10), 830-840.

Chorus, C. G., Arentze, T. A., & Timmermans, H. J. (2008). A random regret-minimization model of travel choice. *Transportation Research Part B: Methodological, 42*(1), 1-18.

Chu, K. K., & Chapleau, R. (2010). Augmenting transit trip characterization and travel behavior comprehension. *Transportation Research Record: Journal of the Transportation Research Board, 2183*(1), 29-40.

Chu, K., & Chapleau, R. (2007). Imputation techniques for missing fields and implausible values in public transit smart card data. *11th World Conference on Transport Research.*

Chu, K., & Chapleau, R. (2008). Enriching Archived Smart Card Transation Data for Transit Demand Modelling. *Transportation Research Record*, 63-72.

Cui, A., Wilson, N., & Attanucci, J. (2006). *Bus Passenger Origin-Destination Matrix Estimation Using Automated Data Collection Systems.* Cambridge: Massachusetts Institute of Technology.

Deng, Z., & Ji, M. (2010). Deriving rules for trip purpose identification from GPS travel survey data and land use data: A machine learning approach. *Traffic and Transportation Studies*, 768-777.

Department for Transport. (2014). *Transport Analysis Guidance unit M3.1 Highway Assignment modelling.* Londen: Department for Transport.

Devillaine, F., Munizaga, M. A., & Trépanier, M. (2012). Detection of activities of public transport users by analyzing smart card data. *Transportation Research Record: Journal of the Transportation Research Board, 2276*(1), 48-55.

Dinant, J. M., & Keuleers, E. (2004). Data protection: Multi-application smart cards: the use of global unique identifiers for cross-profiling purposes-Part II: towards a privacy enhancing smart card engineering. *Computer Law & Security Review, 20*(1), 22-28.

Djukic, T., Hoogendoorn, S., & Lint, H. V. (2013). Reliability Assessment of Dynamic OD Estimation Methods Based on Structural Similarity Index. *Transportation Research Board 92nd Annual Meeting* (pp. 13-4851). Washington DC: Transportation Research Board.

Farzin, J. M. (2008). Constructing an automated bus origin-destination matrix using farecard and global positioning system data in Sao Paulo, Brazil. *TRB Annual Meeting CD-ROM.*

Goeverden, C. v. (n.d.). *Multimodality in the Netherlands 2004 - 2009.* Delft: KIM.

Gong, L., Morikawa, T., Yamamoto, T., & Sato, H. (2014). Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies. *Procedia-Social and Behavioral Sciences, 138*, 557-565.

Gordon, J. B., Koutsopoulos, H. N., Wilson, N. H., & Attanucci, J. P. (2013). Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle Location Data. *Transportation Research Record: Journal of the Transportation Research Board, 2343*(1), 17-24.

Hofmann, M., & O'Mahony, M. (2005). Transfer journey identification and analyses from electronic fare collection data. *Intelligent Transportation Systems* (pp. 34-39). IEEE.

Hofmann, M., Wilson, S. P., & White, P. (2009). Automated identification of linked trips at trip level using electronic fare collection data. *TRB Annual Meeting CD-ROM.*

Jang, W. (2010). Travel time and transfer analysis using transit smart card data. *Transportation Research Record: Journal of the Transportation Research Board , 2144*(1), 142-149.

Jun, C., & Dongyuan, Y. (2013). Estimating Smart Card Commuters Origin-Destination Distribution Based on APTS Data. *Journal of Transportation Systems Engineering and Information Technology, 13*(4), 47-53.

Kieft, S., Herder, J., & Pieters, M. (2013). Openbaar Vervoer Matrices in VENOM. *Colloquium Vervoersplanologisch Speurwerk (CVS).* Rotterdam.

Kieft, S., Linden, T. v., Bedem, J. v., & Scholten, M. (2014a). *VENOM2013 Basismatrices 2010.* Amsterdam: City Region of Amsterdam (SRA).

Kieft, S., Linden, v. d., Bedem, v. d., & Scholten, M. (2014b). *VENOM2013 Basisprognoses 2030.* Amsterdam: Stadsregio Amsterdam (SRA).

Kim, K. (2014). Discrepancy Analysis of Activity Sequences. *Transportation Research Record: Journal of the Transportation Research Board, 2413*(1), 24-33.

Kim, K., Oh, K., Lee, Y. K., Kim, S., & Jung, J. Y. (2014). An analysis on movement patterns between zones using smart card data in subway networks. *International Journal of Geographical Information Science, 28*(9), 1781-1801.

Kuhlman, W. (2014). *Onderzoek verrijking OV chipkaart data met informatie uit de WROOV onderzoeken.* Zoetermeer: Panteia.

Kusakabe, T., & Asakura, Y. (2014). Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies, 46*, 179-191.

Lee, S. G., & Hickman, M. (2013). Are Transit Trips Symmetrical in Time and Space? *Transportation Research Record: Journal of the Transportation Research Board, 2382*(1), 173-180.

Lee, S. G., & Hickman, M. (2014). Trip purpose inference using automated fare collection data. *Public Transport, 6*(1-2), 1-20.

Lee, S., Hickman, M., & Tong, D. (2012). Stop Aggregation Model: Development and application. *Transportation Research Record: Journal of the Transportation Research Board , 2276*(1), 38-47.

Lianfu, Z., Shuzhi, Z., Yonggang, Z., & Ziyin, Z. (2007). Study on the method of constructing bus stops OD matrix based on IC card data. *Wireless Communications, Networking and Mobile Computing*.

Liao, C. F., & Liu, H. X. (2010). Development of Data-Processing Framework for Transit Performance Analysis. *Transportation Research Record: Journal of the Transportation Research Board, 2143*(1), 34-43.

Ma, L. C., Banerjee, P., Lai, J. H., & Shroff, R. H. (2008). Diffusion of the 'Octopus' Smart Card E-payment System: A Business and Technology Alignment Perspective. *International Journal of Business and Information, 3*(1).

Ma, X.-l., Wu, Y.-J., Wang, Y.-h., Chen, F., & Liu, J.-f. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C, 36*(1), 1-12.

McGowen, P., & McNally, M. (2007). Evaluating the potential to predict activity types from GPS and GIS data. Washington: Transportation Research Board 86th Annual Meeting.

Montini, L., Rieser-Schüssler, N., Horni, A., & Axhausen, K. W. (2014). Trip Purpose Identification from GPS Tracks . *Transportation Research Record: Journal of the Transportation Research Board, 2405*(1), 16-23.

Morency, C., Trepanier, M., & Agard, B. (2007). Measuring transit use variability with smart-card data. *Transport Policy, 14*(3), 193-203.

Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies, 24*, 9-18.

Munizaga, M. A., Devillaine, F., Navarrete, C., & Silva, D. (2014). Validating travel behavior estimated from smartcard data. *Transportation Research Part C: Emerging Technologies, 44*, 70-79.

Nassir, N., Khani, A., Lee, S. G., Noh, H., & Hickman, M. (2011). Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system. *Transportation Research Record: Journal of the Transportation Research Board, 2263*(1), 140-150.

Nes, R. v. (2002). *Design of multimodal transport networks: A hierarchical approach.* Delft: Delft University Press.

Nijënstein, S., & Bussink, B. (2014). Verkenning kwaliteitsverbetering OV met multimodale OV-chipkaart data (in Dutch). Eindhoven: CVS.

Oort, N. v., Drost, M., & Brand, T. (2014). Betere OV prognoses met anonieme OV-chipkaartdata (in Dutch). Eindhoven: CVS.

Ordóñez, S. A., & Erath, A. (2013). Estimating Dynamic Workplace Capacities by Means of Public Transport Smart Card Data and Household Travel Survey in Singapore. *Transportation Research Record: Journal of the Transportation Research Board, 2344*, 20-30.

Ortúzar, J. d., & Willumsen, L. G. (2011). *Modelling transport - fourth edition.* Chichester: Wiley.

Pelletier, M., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C, 19*(4), 557-568.

Pollard, T., Taylor, N., & van Vuren, T. (2013). Comparing the Quality of OD Matrices in Time and Between Data Sources. Frankfurt: The Association for European Transport .

Reddy, A., Lu, A., Kumar, S., Bashmakov, V., & Rudenko, S. (2009). Entry-Only Automated Fare-Collection System Data Used to Infer Ridership,Rider Destinations, Unlinked Trips, and Passenger Miles. *Transportation Research Record: Journal of the Transportation Research Board, 2110*(1), 128-136.

Rijkswaterstaat. (2012). *Documentatie Groeimodel 2011 Deel 1.* Rijkswaterstaat.

Robinson, S., Narayanan, B., Toh, N., & Pereira, F. (2014). Methods for pre-processing smartcard data to improve data quality. *Transportation Research Part C: Emerging Technologies, 49*(1), 43-58.

Seaborn, C., Attanucci, J., & Wilson, N. H. (2009). Analyzing multimodal public transport journeys in London with smart card fare payment data. *Transportation Research Record: Journal of the Transportation Research Board, 2121*(1), 55-62.

Shen, L., & Stopher, P. R. (2013). A process for trip purpose imputation from Global Positioning System data. *Transportation Research Part C: Emerging Technologies, 36*, 261-267.

Stopher, P., FitzGerald, C., & Zhang, J. (2008). Search for a global positioning system device to measure person travel. *Transportation Research Part C: Emerging Technologies, 16*(3), 350-369.

Train, K. E. (2009). *Discrete choice methods with simulation* (2nd ed.). New York: Cambridge university press.

Translink. (2015). *Jaaroverzicht Translink 2014.* Utrecht.

Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. *Journal of Intelligent Transportation Systems: Technology, Planning and Operations, 11*(1), 1-14.

Utsunomiya, M., Attanucci, J., & Wilson, N. (2006). Potential uses of transit smart card registration and transaction data to improve transit planning. *TRB Annual Meeting CD-ROM.*

Wang, W., Attanucci, J. P., & Wilson, N. H. (2011). Bus Passenger Origin-Destination Estimation and related analyses using automated data collection systems. *Journal of Public Transportation, 14*(4).

Wolf, J., Guensler, R., & Bachman, W. (2001). Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record: Journal of the Transportation Research Board, 1768*(1), 125-134.

Yue, Y., Lan, T., Yeh, A. G., & Li, Q. Q. (2014). Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. *Travel Behaviour and Society, 1*(2), 69-78.

Zhao, J., Rahbee, A., & Wilson, N. H. (2007). Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems. *Computer-Aided Civil and Infrastructure Engineering, 22*(5), 376-387.

Zhou, J., Murphy, E., & Long, Y. (2014). Commuting efficiency in the Beijing metropolitan area: An exploration combining smartcard and travel survey data. *Journal of Transport Geography, 41*(1), 175-183.