Aggregated GSM data in Origin Destination studies

Final report

Thesis Project
Aggregated GSM data in Origin Destination studies

Final Report

Thesis project

This document is submitted as part of the thesis for the Transport and Planning department, Civil Engineering and Geosciences faculty of Technical University of Delft.

Author: Alexandros E. Papacharalampous - 4240219 ■ +31 (0) 626 177107 ■ alexpap@live.com

Assessment committee: prof. dr. ir. E. de Romph CiTG / TU Delft
                     dr. ir. A.J. Pel CiTG / TU Delft
                     ing. A. P. Knol Dat.Mobility B.V.
                     dr. J.A. Hidders EWI / TU Delft
                     ir. P.B.L. Wiggenraad CiTG / TU Delft

Keywords: transport planning, big data, origin-destination matrices, mobile phone data

Date of completion: Wednesday, 19 November 2014

Document Statistics: 36,086 words in 86 pages (46 Figures and 24 Tables)

This document is optimized for web view.
A request for the appendices can be sent to the author.
Preface

This thesis is submitted as part of the requirements for obtaining the degree of Master of Science of the track Transport & Planning, faculty of Civil Engineering and Geosciences, Technical University of Delft. The thesis was completed at Delft University of Technology and at the company Dat.Mobility in Deventer. This report covers the subject of obtaining Origin-Destination information from the viewDat database provided by Dat.Mobility and from aggregated GSM datasets in general.

This thesis is intended for transportation engineers that use commercially available datasets which are usually in an aggregated form and for data handlers who create these datasets. It provides 3 methodologies that could be used for extracting origin destination information from such datasets, makes suggestions on their most useful specification and their limitations, deliberates on potential uses and discusses the observations made by using aggregated GSM datasets.

I would like to thank Dat.Mobility B.V. for allowing me to work at their project and for using their database. I would extend my gratitude to all the people of Dat.Mobility and Goudappel Coffeng for their hospitality and assistance that made this work possible. I would like to thank in particular ir. K. Zanema, E. Bernards, B. van der Griendt, J. Herder and N. Aardoom for their support in making this research work possible. Furthermore, I want to thank the assessment committee for their reviews and comments; their guidance assisted me in extending my skills and to reach an outcome that I am proud of. Finally, I would like to thank my parents, E. Papacharalampous and V. Tzitzikopoulou, for their strong support in the process of this work and in adjusting to the new environment that this thesis was completed in.

In the realm of Big Data, the improvements are faster than anticipated. New data and data sources are coming to light every day and their analysis is becoming more pivotal than ever. That is because: “You can have data without information, but you cannot have information without data”, by Daniels Keys Moran.

Sincerely,
Alexandros E. Papacharalampous
Delft, November 2014
Summary

Problems in society are becoming increasingly complicated. On the other hand, the information available is also increasing. The increase in information leads to new possibilities for solving these problems and creates the background for improvements in the existing services and technologies. The plethora of information requires techniques and innovations to assist in its usage meaning to solve questions such as where, when, how and why should this information be used.

Mobility planning and management is one of the areas on which Big Data had a great impact. That is because transportation planning is governed by many variables; it evolves with small changes of the circumstances. A better understanding in the movements of network’s users will go a long way towards improving the efficiency of the planning process. For that purpose, GSM data can be employed and a comprehensive analysis of their potential can crucially improve the transport planning process and the accuracy of its results.

Nevertheless, as every new technology, Big Data for transportation are at their infancy. Multiple specifications and types of datasets exist. There has not been a consistent commercially used database specification that could be used in real-time planning processes and that will impact on actual road design. Researchers focus mostly on creating algorithms that can extract the required information for such purposes. This thesis has the opportunity to work with a commercial dataset that has the potential and ability to be used in practice. The aim of this thesis, is to extend that potential by proposing and deriving new uses but more predominantly, on extracting Origin-Destination information.

Goal of the Research and contribution

The research aims to derive Origin-Destination matrices from the database provided by Dat.Mobility B.V. The database is created by Mezuro B.V. which analyses the raw data that it receives from a large mobile network operator. The dataset at the level of the mobile network operator is comprised of traces of users that are anonymized by hiding the telephone number. Mezuro analyses the data and delivers a dataset to Dat.Mobility. This dataset is an aggregated dataset that contains information on how many persons where located per region for specific time intervals. The dataset that Dat.Mobility receives also contains information on the home location of users. However, due to the aggregated nature of the database, it is impossible to relate persons with home locations.

GSM datasets have not been around the transportation sector for long. It is an innovative and low-cost idea to obtain information for the movements of users in comparison to the household surveys which are expensive while the sample size is very small. However, from the side of privacy concerns the thin line between traces and aggregated statistics is not usually crossed by companies which commercialize GSM data. In other words, companies prefer to give aggregated datasets that have a reduced capability to extract information compared to the raw data. Therefore, the question that arises is how can engineers efficiently derive as much information as possible from these datasets.

The goal of this thesis is to answer a part of this question which relates to one of the most useful pieces of information for transportation modelling, Origin-Destination matrices. To achieve that, the existing Origin-Destination matrices’ configuration will be explored. That way, a proposal for the
specification of the data of Dat.Mobility can be formulated. It is also required in order to assess the effectiveness of the existing specification.

The data are moreover analysed concerning their limitations and inconsistencies. This is pursued in order to comprehend the limitations of the derived Origin-Destination matrix and to propose improvements in the dataset or methods to solve existing problems. Finally, a side goal of the thesis is to propose additional uses of the database, besides the derivation of Origin-Destination matrices.

Research Methodology

The research is comprised of three parts. First part is the data analysis where the number of residents in total and per region is checked for consistency. The determination of the home locations should be the most accurate since it is based on a whole month of data. Additionally, Mezuro divided the users in three categories based on the number of times they visit an area. This interesting aspect is analysed in correlation to trip purpose, i.e., work, shopping. Finally, this part analyses in extent the limitations of the database which are the foundations for the next part.

The second part builds up on the limitations that were discovered before to create assumptions necessary to create the Origin-Destination matrix. Based on the assumptions, a method is devised that can provide the morning peak Origin-Destination matrix and that can be easily reproduced per day. The derived Origin-Destination matrix is then checked for accuracy by comparing it to observed and modelled, yet extensively used, values.

The third and final part of the analysis is targeted to aggregate GSM datasets in general. The purpose is to create methods that can be used to extract Origin-Destination matrices from aggregate GSM datasets. The methods also look into the specification of the dataset always attempting to find the simplest one that can be used for the aforementioned purpose.

Results

It was seen that the number of residents varies monthly at an average of 100,000 residents. This number becomes more stable after June while months before June were found to be inconsistent for various reasons. In essence, the algorithm that Mezuro uses for determining residents was at its infant steps in the first months of 2013. Thus, the degree of variation that was seen is expected and a part of it can be attributed to the robustness of the estimation procedure. Nevertheless, the variation of residents requires further research since it may reveal causal mechanisms for population mobility at a deeper level that is not accounted for in classical modelling.

Furthermore, the number of residents was much lower than the number of active users that appeared in the dataset. It is unclear who these excess active users are. It could be that a number of individuals that appeared in the dataset are actually people that do not live in the Netherlands or, more importantly that they are people who own more than one mobile phone. If the latter is the case, then different mobile phones may account for different movements of people. For example, a person might use a mobile phone for business purposes and another for personal purposes. One of those phones might be deactivated in the night hours and it will not account as a resident. It is important that the phone activity is researched in order to determine such occurrences. Phone
activity is also important for determining the size of the sample and more specifically, for scaling the amount of users that are active to the size of the Mobile network operator sample.

Mezuro uses three categories according to how many times a person visited an area. These categories were found to correspond to the trip purposes, e.g., the frequent visitor category was corresponding to business trips and trips to work. However, none of these categories is uniquely corresponding to a single purpose. The findings show that this relationship could be further exploited to derive purpose. This would require additional rules to divide into categories besides the number of times that a person visits an area.

Deriving Origin-Destination matrices from the database is not a straightforward procedure because the dataset lacks the connection between origin and activity location. Although the HPP dataset contains spatial connections that could be mistaken for trips, the definition of the dataset does not allow the direct use of the observations. First, the HPP dataset is in periods which require disaggregation to hours, the time interval used in Origin-Destination matrices. Therefore, the dataset contains trips that span for more than an hour and also contains additional trips of the same user, named in this thesis ‘excess’ trips. The definition of the dataset also allows multiple counting of the same individual in different regions. The multiple counts were estimated for a 2-hour interval to be 21.5% of the total counts. The ‘excess’ trips, are estimated to be 42.5% of the total HPP counts in the same interval.

The locations where individuals participated in activities could be found from the DPU dataset because this dataset logs the location were the individual was mostly seen. This is the best equivalent of activity that can be found in the dataset. Consequently, the algorithm that was created, uses the DPU dataset’s counts of 10:00a.m. as destinations and the HPP dataset’s counts of 10:00a.m.–12:00p.m. as origins and as OD pairs’ values. The time interval that was used had to be as close to morning peak as possible because the only origin in the HPP dataset is the home location thus, only trips from home can be analysed. Additionally, using the 10-11h time interval of the DPU dataset ensures that the travellers of the morning peak, who travelled in the interval 7-9h and in the peak shoulder which is 9-10h, have reached their destination. The values of total origins and the OD pair’s values were equilibrated to the destinations. As a final point, it should be noted that the algorithm created is not computationally intensive and can be used to estimate an OD matrix per day in less than 5 minutes as it was measured on an average machine.

The accuracy of the matrix was determined by the accuracy of the assumptions required for creating it. For that purpose, the trip length distribution was compared to an observed distribution from the Monitoringstool 2009, the sparseness of the matrix was compared to the sparseness of the LMS and specific regions were analysed to determine for specific (or otherwise) inconsistencies. The matrix fared well in almost all accuracy tests. However, some tests yielded some questions, such as why is the ratio of origins over destinations always higher for the derived OD matrix compared to LMS. These questions were catalogued and are proposed as further research. Additionally, they allowed the formulation of a proposal to improve the HPP dataset and that will improve the accuracy of the final OD.

Two methods were also formulated and discussed that could assist other endeavours of deriving origin-destination information from aggregated GSM datasets. One method is data focused and exploits the fusion of two datasets that are similar to the viewDat database but simpler in the sense that less analysis of the raw data is required. The second method is a probabilistic and data intensive approach. Nevertheless, most calculations could be accomplished offline which makes it valuable for dynamic applications.
Conclusions and Findings

This research used a newly developed method of comparing matrices, the structural similarity index which was found extremely useful given the context of this thesis. It is the author’s belief that the tool is very powerful for comparing matrices and very useful since its application is not complicated. However, there is no indication of the span of possibilities for this tool to be used in practice. Additionally, this thesis found that there can be mistakes in its application and its estimation, for example, when the compared matrices have low variances and a low covariance. For that purpose, some remarks are made concerning its application and a proposed change in the formulation is given when the index is used for comparing matrices. Additionally, this thesis found that the index can be used for comparing two matrices for differences in sparseness, by replacing the non-zero elements with 1.

Furthermore, determining the accuracy of the derived Origin-Destination matrix required the comparison of two matrices that had a different spatial specification. No specification was coarser or more detailed than the other. In other words, there were areas that were more detailed in one specification and coarser for the other and vice versa for other areas. The comparison of the two matrices was achieved by referencing both matrices to a common spatial specification. This was a simple rectangular specification where the sizes of the rectangles were chosen so that the maximum number of rectangles contained at least one centroid from either map specifications. This simple algorithm could be improved in various ways but provides the basis for comparing maps of different specifications.

It was also found from the literature review that there exist two different definitions for aggregated GSM data. In this research, there was the opportunity to work with both. A SWOT analysis was done in order to summarize and compare the aspects of these specifications to each other and to disaggregate GSM data.

Aggregate GSM data proved to have great potential for a number of transport planning applications. In essence, aggregate GSM data are population counters and could have great potential to replace count data especially in regions where the cost of installing counters is of major interest. This could be achieved by gathering data from specific antennas that are close or mainly serve the traffic in a highway or road. This shows that the use of aggregate GSM data can extend further than creating an OD matrix.

The most important conclusion of this research work is that deriving Origin-Destination matrices requires only a small amount of information. Origin-Destination matrices can be extracted by fusing two datasets where one shows the most active location and the other contains information about the home location. Both datasets could be created by using simple spatiotemporal rules. The added value of this result is that the process of creating an Origin-Destination matrix per day is short and can be reproduced easily while the potential of using this process in dynamic modelling is extremely high.
Table of Contents

1. Introduction .................................................. 1
   1.1 Scope of research ......................................... 2
   1.1.1 Research questions ................................... 2
   1.1.2 Research approach ..................................... 3
   1.2 Structure of report ........................................ 4

2. Literature Review ............................................. 5
   2.1 User localization with GSM data ......................... 5
   2.1.1 The GSM physical network and users’ location estimation 5
   2.1.2 Types of GSM data for positioning users ............ 6
   2.1.3 Location estimation techniques in practice .......... 7
   2.2 Origin-Destination matrices and the modelling process 8
   2.3 Origin-Destination matrices from GSM Data ............ 10
   2.4 Other mobility studies using GSM data ................. 12
   2.5 The understudied database of Dat.Mobility B.V. ....... 13
   2.5.1 The Mezuro Data ..................................... 13
   2.5.2 The viewDat database ................................ 14
   2.5.3 A preliminary view on the limitations of the viewDat database 15
   2.6 Setting the Initial Assumptions ......................... 16
   2.7 Conclusions ............................................. 17

3. Data analysis .................................................. 18
   3.1 Introductory Remarks ..................................... 18
   3.2 Analysing the total number of residents .............. 18
   3.3 Analysing the number of residents per location ...... 21
   3.4 Analysing the visitors’ group of the viewDat database 23
   3.4.1 Examining the frequent visitors of the 0-24h period 23
   3.4.2 Comparison of the Trip Purpose graphs with viewDat data variations 24
   3.5 Limitations and issues of the dataset .................. 28
   3.6 Conclusions ............................................. 34
   3.6.1 The problem of the sample’s size ................... 35
4. **Acquiring Origin-Destination information from the viewDat database**
   4.1 Initial Remarks
   4.2 List of Assumptions
   4.3 The process for creating the OD matrix
   4.3.1 The equilibration procedure of OD pairs to total origins/destinations
   4.3.2 A solution to the ‘Less than 15’ problem
   4.4 Accuracy of the generated OD matrix
   4.4.1 Comparing the Trip Length Distributions
   4.4.2 Sparseness of the matrix
   4.4.3 Examining specific regions
   4.5 Conclusions about the accuracy of the assumptions
   4.6 Recommendations
   4.6.1 An improved specification of the HPP dataset
   4.6.2 The SSIM with low variance and/or covariance

5. **Additional methods for extracting OD matrices from aggregated data**
   5.1 Initial remarks on aggregated GSM datasets
   5.2 Possibilities for deriving OD information from aggregated GSM data
   5.2.1 A data-focused method for excluding ‘passers-by’
   5.2.2 A probabilistic approach for excluding ‘passers-by’
   5.3 Conclusions

6. **Conclusions and Answers to the Research Questions**
   6.1 Answers to research questions
   6.2 Conclusions
   6.2.1 Conclusions from the process of the research
   6.2.2 Extracting OD information from aggregated GSM datasets
   6.2.3 A comparison of different specifications of GSM data

7. **Recommendations and Future Research**
   7.1 Recommendations
   7.1.1 Using the viewDat database for calibration/validation
   7.1.2 Optimal specification of the viewDat database to extract OD matrices
   7.2 A recommendation for an improved methodology for applying the SSIM
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Future research</td>
<td>74</td>
</tr>
<tr>
<td>7.3.1 On the viewDat database and Mezuro data</td>
<td>74</td>
</tr>
<tr>
<td>7.3.2 On aggregated GSM datasets</td>
<td>75</td>
</tr>
<tr>
<td>Closing Remarks</td>
<td>76</td>
</tr>
<tr>
<td>Bibliography</td>
<td>78</td>
</tr>
<tr>
<td>Lessons Learned – A short reflection on the research process</td>
<td>84</td>
</tr>
</tbody>
</table>
List of Figures

Figure 1: Index of the search term "Big Data" .......................................................................................... 2
Figure 2: Cell towers around Deventer, Overijssel ................................................................................. 6
Figure 3: Positioning of device using event-driven data or CDR ......................................................... 7
Figure 4: Positioning of device using network-driven data ................................................................. 7
Figure 5: Fundamental procedure for predicting a year using the LMS model ................................ 10
Figure 6: Motifs from household surveys and GSM data ................................................................. 12
Figure 7: Change in data type from the original data source until creation of the viewDat database ............................................................................................................................ 14
Figure 8: Average number of residents for all the Netherlands per month ....................................... 19
Figure 9: Average number of active users for all the Netherlands per month .................................. 20
Figure 10: The difference of residents from GSM to actual residents of regions of the Netherlands ................................................................................................................................. 21
Figure 11: Residents from GSM data for July versus residents from NRM data .................................. 22
Figure 12: Time series model for July 2013 and the forecast for August compared to the data of August ........................................................................................................................................ 24
Figure 13: Variation of frequent visitors within year 2013 and the trend of the time series ............ 24
Figure 14: Weekly variation of the frequent visitors for September and October ............................... 25
Figure 15: Weekly variation of the regular visitors for September and October ................................ 25
Figure 16: Weekly variation of the incidental visitors for September and October ........................... 25
Figure 17: Trip distribution per person per purpose and per week day .............................................. 26
Figure 18: Trip length distribution and Trip distribution per purpose ............................................... 27
Figure 19: Antennas that a mobile device at TU Delft can connect to ............................................... 30
Figure 20: Routes between Utrecht and Deventer ................................................................................... 31
Figure 21: Multiple counts of origin Pt.1 ............................................................................................... 31
Figure 22: Multiple counts of origin Pt.2 ............................................................................................... 32
Figure 23: Multiple counts of origin Pt.3 ............................................................................................. 32
Figure 24: Differences between BAG and LMS zoning system specification for Terschelling .............................................................................................................................................. 33
Figure 25: Differences between BAG and LMS zoning system specification for Amsterdam and Almere ........................................................................................................................................... 33
Figure 26: Comparison of the trip length distributions of the Monitoringstool 2009 and the OD from the viewDat database ........................................................................................................................................... 43
List of Tables

Table 1: Type of mobile phone location data................................................................. 13
Table 2: The event-based log that constitutes the Mezuro data........................................... 14
Table 3: Summary of limitations of viewDat database......................................................... 15
Table 4: Average number of residents per month ............................................................. 19
Table 5: Amount of outliers per month .............................................................................. 20
Table 6: Coefficients of Linear Regression........................................................................ 23
Table 7: Possible relation of trip purpose and type of visitor
   based only on the shape of the variation......................................................................... 27
Table 8: Summary of issues with the databases that damage the quality of the OD matrix ....... 28
Table 9: Sparseness of the LMS matrices........................................................................... 46
Table 10: Calculated origins and destinations for selected regions....................................... 48
Table 11: Ratio of origins over destinations for viewDat database and LMS......................... 48
Table 12: Origins and destinations per the selected regions
   over total trips for the viewDat database and LMS....................................................... 48
Table 13: Trip ratios to Amsterdam and The Hague with the selected regions as origins....... 49
Table 14: Cumulative table of the limitations and possible solutions.................................. 64
Table 15: SWOT analysis on disaggregate GSM data.......................................................... 68
Table 16: SWOT analysis on viewDat data......................................................................... 68
Table 17: SWOT analysis on framing data......................................................................... 69
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAG</td>
<td>Register of addresses and Buildings (Basisregistraties Addressen en Gebouwen)</td>
</tr>
<tr>
<td>CDR</td>
<td>Call Detail Record</td>
</tr>
<tr>
<td>DPU</td>
<td>Density Per Hour (Dichtheid Per Uur)</td>
</tr>
<tr>
<td>(E)TLD</td>
<td>(Empirical) Trip Length Distribution</td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobile communications</td>
</tr>
<tr>
<td>HPP</td>
<td>Origin Per Period (Dichtheid Per Periode)</td>
</tr>
<tr>
<td>LMS</td>
<td>National Model System (Landelijk Model Systeem)</td>
</tr>
<tr>
<td>OD Matrix</td>
<td>Origin-Destination Matrix</td>
</tr>
<tr>
<td>PT</td>
<td>Public Transport</td>
</tr>
<tr>
<td>TAZ</td>
<td>Traffic Analysis Zones</td>
</tr>
</tbody>
</table>
1. Introduction

Big Data have the potential to become an essential tool in transportation planning and modelling. They can provide a richer dataset of the movements of individuals and an alternative to the costly in money and time, traditional data acquisition techniques, namely household surveys and in-situ vehicle counts using induction loops. Companies in the transport sector have since some time danced around the idea of obtaining big data, for example, by requesting the user to opt-in passive data sending which is the case with floating car data. The efficient estimation of the movements, in real time and for strategic planning is huge.

GSM data have a great potential to meet the goals of planners and modellers. The database's size is unprecedented and most importantly, GSM data are readily available every hour of a day in a continuous form. They include information about all travel modes and trip purposes of travellers. However, the raw form of the data requires the derivation of methodologies for positioning users in the geo-spatial plane, for determining mode, speed, travel time and other transportation variables (for a review of the existing mobile data technologies see: Rose, 2007; Caceres, et al., 2008; Wolf, et al., 2014a).

GSM data can be disaggregate and aggregate (Calabrese, 2011). The definition of disaggregated GSM data is clear; individual GSM data allow the creation of a single user's route in the spatial plane. That is the reason that these data are referred to as individual GSM data. The definition of the aggregated GSM data however is not clear. Wolf et al. (2014a) especially reviewed formats of aggregated and disaggregated GPS datasets from cellular devices that are obtained from private firms. They mention that the contents of files of aggregated datasets include origin zone, destination, trip type, e.g., work, education, and number of trips. This definition strays away from the one stated by Calabrese (2011) which are aggregated phone statistics at the antenna or another spatial level.

In the private company level, the availability and disaggregation of data is a major concern because it is related to privacy issues which are protected by national and international laws. Private companies prefer to not offer disaggregate datasets not only to abide to legality but also to protect their public persona since a number of users are concerned with the tracking capabilities of new technology. This inquietude of users was reviewed by Pierce et al. (2011). Private companies prefer to maintain data management without sharing their methodologies and prefer to protect their data sources which essentially are their individual customers (Wolf, et al., 2014a).
The turning point for GSM data studies should be user awareness on the data they transmit from their personal devices. Boyles et al. (2012) conducted a phone interview survey on 2,254 individuals and reported that 57% of app users either deleted or did not install an app to begin with, because they had privacy concerns over the sharing of their personal information that this app requested. This shows how sensitive users are to sharing their personal information with another and more importantly a corporate entity. With the interest on Big Data being on the rise (see Figure 1), questions and inquiries on the data specification and the methods they are obtained and stored, are bound to increase.

![Figure 1: Index of the search term "Big Data", Source: https://www.google.com/trends/explore#q=big%20data&date=1%2F2007%2085m&cmpt=q](https://www.google.com/trends/explore#q=big%20data&date=1%2F2007%2085m&cmpt=q)

The future of Big Data is uncertain and dependent to many political and societal pressures. Private firms seem to be aware of this fact and are self-regulated to not share information on the individual level. The protection of anonymity is a big component of companies’ public image and sometimes is used as a trade product. For example, consider the contemporary social network named ‘Ello’. ‘Ello’ is advertised as anonymous in contrast of other social networks such as ‘Facebook’ where ‘Every post you share, every friend you make, and every link you follow is tracked, recorded, and converted into data.’ (Source: [https://ello.co/wtf/post/manifesto](https://ello.co/wtf/post/manifesto)). The network is ‘invite-only’ and became quite popular very fast to the point that an invitation for joining it was sold at Ebay™ for up to 100$ (Woolf, 2014).

But the richness of the GSM location datasets cannot be ignored. The valid question that should be posed is how the required information for transport planning and modelling applications can be extracted from simpler, more aggregated datasets or datasets which de facto protect the anonymity of users.

This thesis will be employed with determining the feasibility of extracting Origin-Destination information from aggregated GSM datasets and for deriving methodologies for that purpose. It will also make a comparison of an aggregated GSM dataset to the individual GSM datasets and explore what would be the best specification of an aggregated GSM dataset that has the maximum added value to transportation engineers.

1.1 Scope of research

1.1.1 Research questions

The thesis will focus on estimating origin destination matrices and determine their accuracy by comparison to existing OD matrices or other measured statistical values. It will also attempt to determine what alterations can and should be made to the specification of the datasets in order to improve the OD estimation.
Additionally, in the thesis’ timeframe, new data of completely different specification will become available. These data will be analysed and the results are included in this document.

The main research question is:

*Can Origin-Destination matrices be extracted from aggregated GSM datasets, how can this be achieved and how accurately do they represent reality?*

The side research questions that will support the main research question are:

- What is the spatial and temporal aggregation used in existing transport planning models and in what level of transport analysis can the data be useful?
- What are the limitations and inconsistencies of the data and how can they be solved?
- What are the possible uses of the viewDat database besides deriving Origin Destination information?
- If Origin-Destination matrices cannot be derived from the database, then what additional information is required from the data to achieve this purpose?

1.1.2 Research approach

Establishing whether Origin-Destination matrices can be extracted from the database will not only provide a solution for the specific database to be used ad hoc but will also be useful for extracting origins and destinations from similar databases. But when is a database considered similar? First, any database that is used for transportation planning purposes has to be in line with the requirements of transportation modelling. Therefore, determining the appropriate spatial and temporal specification will be of great use for the data handlers who deliver the data to the transport engineers. By doing so, the entry level for the data can be determined, i.e., strategic planning, national level. This is what the first research question aims to solve.

The second research question is more focused on the existing dataset. Understanding the limitations of the existing dataset is indeed helpful for devising and applying methodologies to derive origin-destination matrices, i.e., for establishing the required assumptions that lead to such methodologies. Listing and discussing the limitations is also important for informing other companies and researchers of what issues they might encounter when working with aggregated datasets and how they can avoid them.

The third research question is an extension of the main research question. It extends the view from Origin-Destination matrices and determines what other opportunities can be pursued using aggregate GSM data. Furthermore, this question relates to not only the transport planner but the data handler as well, discussing opportunities of the data in the transport planning context.

The final side-research question is meant to support the main research question in the case that Origin-Destination matrices cannot be extracted. It also extends the main research question by essentially asking how the accuracy of the matrices that are derived, can be improved.

The main goal is pursued using a database that is constituted of more than one year of data divided into two subsets that contain data per hour and per period.

The literature review and the data analysis set the foreground for creating OD matrices from the datasets. Due to the fact that this thesis focuses on different data than the common literature but that are adjacent to the analysis of mobile phone data, the literature review assists in setting the borders of good and bad practice. Both the literature review and the data analysis assist in proposing an improved specification for the datasets.

The derived Origin-Destination matrices are always compared to measured/observed statistical values or other established modelled values (LMS matrices) that are commonly used in practice. Comparing to modelled values creates the problem of not knowing which value is correct. If both the dataset and the modelled values seem to be equal, then it is a good indication that the derived
Origin-Destination matrices are correct. Or that they are both wrong. However, the fact that the LMS matrices are commonly used in practice, make them a reference point for the derived matrices.

1.2 Structure of report

There are 7 chapters in the thesis including the present. The 2nd chapter contains the literature review which focuses on previous work using mobile phone data and explains the specification and other characteristics of the datasets. The 3rd chapter analyses the datasets in depth, determining temporal variations and attempting to find and resolve inconsistencies. The analysis concerns aspects that are of interest for Origin-Destination estimation such as trip purpose while it lists the limitations that cannot be solved from the side of Dat.Mobility.

The 4th chapter describes the method that was used to create the OD matrix. It also discusses the results of different tests that were done to determine the accuracy of the derived OD matrix. The 5th chapter is providing 2 methods for determining Origins and Destinations from aggregated datasets and also provides the specification that is required for applying the methodologies. The 6th chapter discusses the conclusions and the answers to the research questions The 7th chapter includes the recommendations concerning the usability and improvement of the datasets and gives indications for future research.

There are 4 appendices created during the research progress. The first has information on the architecture of the GSM dataset that could be useful for the intrigued reader. The second contains detailed information on the specification of the viewDat database that is used throughout this research work. The third contains information about the framing data, a new type of dataset that is available from Mezuro and that contains directly origin-destination information. Finally, the fourth discusses an alternative concept for deriving trip purpose from aggregated datasets.
2. Literature Review

2.1 User localization with GSM data

This chapter analyses, as a prelude to more specific subjects such as Origin-Destination matrix estimation, how the user is located in the spatial place using GSM data. Inherently, all devices/individuals are merely datum points which would hold no meaning for transportation engineering without the connection to a geographical system. This chapter analyses basic aspects of the GSM network required to understand how the positioning of the user takes place which is also put under the microscope further ahead.

It should be noted that GSM data and mobile phone data are used interchangeably in this thesis although the latter is more general. Essentially, GSM data are a part of mobile phone data.

2.1.1 The GSM physical network and users’ location estimation

The GSM physical network, also known as the cellular network, is comprised of antennas/cell towers. Each antenna is part of a cell. A cell is a theoretical geographical area comprised of 1-3 antennas. Practically, each device that is inside a cell is connected to one of the antennas that comprise the cell. Each antenna has a service area which is the distance from the antenna within which a mobile phone is capable to connect to the antenna. The service area of an antenna is the range in distance terms within which the antenna can maintain a stable connection with mobile phones. Each antenna can maintain only a number of connections before overloading. To avoid overloading, mobile phones connect to other antennas that have a smaller load. It can then be reasonably assumed that service areas of antennas overlap. (Kwan, 2012)

The service areas of antennas can reach up to 35km in rural areas and is usually within the range of 0.5-5km within city limits (Kwan, 2012). There also exist antennas with smaller range that service small areas unreachable by the signal of the main antennas such as subway stations. All these antennas cover the complete stretch of the Netherlands.

Counting of device/individuals using GSM technology is as simple as counting the number of times that a user connected to an antenna. When a user makes a phone call, he/she creates a stream of data between his/her phone and an antenna. Information about the location/antenna, type of event
and time are stored within the GSM network structure. When a user initiates a stream of information with an antenna, it is considered a sighting, i.e., a time-stamped location of the person who owns the mobile phone. Since individual information is not available, each antenna simply has a number of sightings. These sightings are better aggregated in areas that are of interest for mobile studies. For example, the city of Delft might be divided in some areas based on population density or accessibility while these areas may contain multiple antennas.

2.1.2 Types of GSM data for positioning users
Mobile phone data can be categorized by accuracy and availability. These are complementary terms for the data collection procedure but contradictory in the sense that better accuracy is generally characterized by less availability. From the telecommunications’ operator’s side, there are two big categories of mobile phone data: event and network based (Calabrese, 2011).

Event-driven data are logs created when an event takes place. This event can be a phone call (only begin of call), use of SMS and MMS, data traffic and roaming information. These logs are time-stamped location data that contain information such as the duration of the call or the size of the SMS and are used by the operators for billing purposes. A type of event-driven data and probably the most popular is the Call Detail Records – CDR. These data are very popular among transportation engineers due to their large amount as well as the availability from the telecommunications’ operator. They are provided in a standardized form and are regularly and thoroughly updated by the operator. However, they are probably the least accurate type of mobile phone data (Costabile, 2010; Smoreda, et al., 2013).

Each time a call is placed, the cell where the call took place is logged. That way the user can be located in the specific cell. The cells can be connected to a topographic location through a “cell” map. Therefore, the created log contains a series of locations that the user has gone through.

Network-driven data are also known as technical network logs. Technical network logs are used for the operation of the network. In addition to the data of the CDR, these logs contain end of calls, handovers and Location Area – LA updates. An LA update occurs when a device crosses from one LA to another while handovers are cell updates when a call is underway. The technical network logs need to be extracted from the MSC unlike the CDR that can be extracted in a readable form from the operator (Smoreda, et al., 2013).

Figure 2: Cell towers around Deventer, Overijssel. Source: http://opensignal.com/
The difference in positioning an individual using event or network driven data, can be substantial. A representation of how a device is located for the two cases of data is shown in Figure 3 and Figure 4. The red dots show the cells to which the user is connected based on the information of each log. The numbers show which events correspond to a location update. The black dotted line shows the actual movement of the device/person and the red continuous line shows the movement of the device/person using the positioning from each type of data.

![Figure 3: Positioning of device using event-driven data or CDR](image1)

![Figure 4: Positioning of device using network-driven data](image2)

### 2.1.3 Location estimation techniques in practice

Aggregated GSM data contain information about the GSM antennas that the user was connected to, not his/her exact location (although there are methodologies that can extract a very precise position of the user as Keij (2013) and Costabile (2010) discuss). Datasets of this type usually contain information about the phone activity. An example is the Erlang data that were made available by Telecom Italia. The MIT SENSEable City Laboratory plotted the city of Rome using these data (Caceres, et al., 2007).
Based on the same data, Reades et al. (2007) researched Rome city’s spatial and temporal dynamics. Their preliminary research required to spread the data (that are sightings of persons/devices) which were aggregated at the antennas over a wider area which corresponded to the antenna’s service area. For that purpose they used an exponential distribution model to determine the value per point. However, such a methodology may not be accurate enough for longer distances and depending on mode availability. In this case, there can be a clear distinction between slow modes and fast modes, i.e., between walking and train, the accuracy of such a methodology is debatable.

Ratti et al. (2006) used a different approach for spreading the sightings over a region. They used the centre of gravity of the region that is serviced by an antenna. Consequently, all activities in that antenna are geo-referenced to the centroid of the service area. It was assumed that each antenna’s coverage area extends to 400-500 m radius and to a $120^\circ$ sector.

The correspondence of traffic zones and cells is a topic that is discussed in research and of importance to this thesis, in order to critically analyse the results of the analysis. Cells of the GSM network are usually smaller and more than traffic zones. Therefore it can be that a Traffic Analysis Zone – TAZ contains a lot of cells and that a cell extends to two or more TAZ. The first case can be easily solved by summing the different cells in the TAZ. The latter case can be avoided by assigning the mobile phones that are connected to a cell that is shared in two TAZ according to the proportion of the cell in each TAZ (Fang, et al., 2014). This concept is identical to the centre-of-gravity of Ratti et al. (2006).

The method of Ratti et al. (2006) and Fang et al. (2014) is not a new concept but closely related to the population-weighted centroid (Wang, 2011) and is of interest for this research. Nevertheless, the assumption made by the authors that the average service area of an antenna is 400-500 m is a very coarse approximation since antennas, especially in urban areas, can reach up to 5km (Kwan, 2012). Costabile (2010) mentions that the Cell ID location technology’s accuracy, which was used for the dataset of Ratti et al. (2006) and is also used in the dataset that this thesis will use, is between 300m-20km. It is the least accurate positioning method while the most accurate positioning technology is Radio-Frequency (RF) Fingerprinting with an accuracy of 10m indoors and outdoors (Costabile, 2010).

### 2.2 Origin-Destination matrices and the modelling process

Origin-Destination – OD matrices, are matrices showing the number of trips made between different points in the spatial system. For strategic transport planning models, which this thesis will deal with, the spatial system is divided in traffic zones called TAZ. Traffic zones are regions basically designed by the transport planner and which are usually consistent with some sort of administrative division.

* Trip is movement of a person from one location to another with the purpose of participating in an activity. The person might participate in two or more consecutive activities in one location but will never do more than one trip between two consecutive activities.

The TAZ centroids are the origins and destinations of trips of users. The detail of traffic zones or the zones’ resolution depends on one design perspective: the kind of trips that will be analysed. Intra-
Zonal trips, which are trips inside a TAZ, do not have a geographical reference. For example, moving from one neighbourhood to another may be useful for the desired level of analysis but moving from one shop to the next door may not be. Usually traffic zones are selected so that the population for each zone is the same and that trips from and to each zone are comparable. The latter is usually achieved by setting a threshold for maximum trips per zone, e.g., for a zone of 1,200 to 3,000 people, 15,000 trips can be a trip threshold (Cambridge Systematics, Inc. and AECOM Consult, 2007). There can be a lot of ways to divide into traffic zones using existing spatial systems, i.e., municipalities or census districts (Bell & Iida, 1997).

The OD matrices are part of the 4-step transport planning model. The 4-step model is a framework for model application which includes a number of mathematical models that are applied consecutively. The model calculates network loads using information about the population such as family size or car ownership. This modelling approach is used to make predictions about the changes in network load when there is a change in the characteristics of the population or of the environment, e.g., what will be the change in flow caused by the creation of office spaces in a region or which roads/highways will have insufficient capacity in the future. In The Netherlands, questions such as the latter are solved primarily by applying the LMS (Landelijk Model Systeem) and NRM (Nieuw Regionaal Model) models.

The general framework of the 4-step model is the following:

1. **Trip Generation**
   In this step, the number of trips that is produced at each zone and attracted by each zone is calculated using mostly socio-economic information. A number of methods can be used to model the number of trips and to predict the number of trips in future years.

2. **Trip distribution**
   In the trip distribution step, the trips are distributed among centroids. In the previous step only the total numbers of arrivals and departures were produced for each centroid but the trips between these centroids was not known. This step aims to calculate the trips from each origin to each destination* and save the results in the Origin-Destination-OD matrix.

3. **Modal Split**
   After determining the trips between each zone, the mode choice of the traveller is examined. That means that trips are assigned to a mode according to the preference of the traveller. Trip distribution and modal split are usually intertwined; the final result is an OD matrix showing trips from a zone to another with each of the examined modes.

4. **Route Assignment**
   The final step is assigning the trips found in the previous steps to the routes of the network which will result to the loading of the network. There are a number of algorithms used to assign the trips to routes such as All-or-Nothing assignment which uses Dijkstra’s algorithm (Dijkstra, 1959) for assigning the trips to shortest route. Other algorithms are Deterministic User Assignment and Stochastic Assignment.

*-Origin is the location from where a user began to travel. It also is a place where the user participated in an activity. Home location is also considered an origin.*

**Destination** is a geographical location where the user participated in an activity. Examples of activities are work, education and shopping among others. There cannot be two consecutive destinations that have the same geographical location. Home can also be a destination.*

One more step can be included above, that is departure time choice. In this step, the model assigns travellers based on the time they depart. Time choice relies on preferred arrival time estimated using free flow time and congestion time. Time of arrival has a utility that is represented at the departure time choice. This step is more useful for dynamic applications and is usually omitted for strategic planning. Nevertheless, strategic models usually make a distinction of the time of day they
are calculated for. For example, the LMS model makes a distinction of morning peak (7:00-9:00 am), evening peak (17:00-19:00 pm) and rest of day.

The LMS model that will be used for comparison a number of times later in the thesis is in fact a sophisticated 4-step model. The main process is the same but the LMS is created to predict future situations. Thus, it can be used to make decisions about road expansion, creation or even pricing policies. This is done by applying the model two times, first with observed values for the present situation and second with predicted values for the future situation. The present situation is called base year and the result of applying the 4-step model are the base year matrices. Presently there are two base years, 2004 and 2010 while the latter has been completed only recently. The predictions made with the base year of 2004 are for the years 2020 and 2030.

![Figure 5: Fundamental procedure for predicting a year using the LMS model](image)


### 2.3 Origin-Destination matrices from GSM Data

The previous subchapter briefly explained the steps for creating OD matrices from statistical information. The 4-step model is the fundamental approach for turning population and mobility statistics, such as number of residents or availability of car in the household, to OD matrices. In essence, the OD matrix that is the outcome of this process is a modelled matrix because it is derived from a modelling process. Creating an OD matrix from GSM data on the other hand is not a modelling process. The OD matrices from GSM data are derived from observations which are the sightings at the antenna level. In the simplest of terms, combining an origin and a destination of a single user will result in an OD path.
The origins of research on OD estimation using individual GSM data should be the pioneering work of White & Wells (2002). Initial literature was mainly focused on estimating travel time (Virtanen, 2002; Karhumaki, 2002) and on determining potential applications of GSM positioning data (Astarita & Florian, 2001).

There are a number of projects where anonymous GSM data were used for positioning vehicles and travellers. For example, the Transport Research Laboratory in the United Kingdom attempted to update OD matrices using mobile phone data (Caceres, et al., 2007). It was concluded in the latter project that updating OD matrices using mobile data is possible. An OD matrix created from mobile phone data was compared to the one that already existed for the area of Kent in United Kingdom showing that OD matrices could be calibrated using GSM data (White & Wells, 2002).

In individual mobile phone data, extracting origins and destinations requires the determination of when the traveller has stopped and when he/she is moving. Studying individuals allows for employing stop and go algorithms. This type of algorithm is probably the most extensively researched in individual CDR data and is very central in transportation and GSM data related studies.

A stop & go algorithm does what its name implies; it determines if the user stopped at a location to participate in an activity or if he/she is moving. For example, consider a user moving towards Utrecht from Deventer. The user places a call just before reaching Utrecht and then 50min later in The Hague. The question is if the user has stopped at Utrecht to participate in activity or not. Assume that the train and car, which are the possible mode alternatives used to travel between the two cities, both require more than 40 min which is true in off-peak hours. The question then becomes if the traveller had time to participate in an activity in Utrecht or any other place in between. In this example it is very ambiguous and will rely in the spatiotemporal rule used to determine whether a user participated in an activity.

It has been found that the required amount of time for a user to have participated in an activity is 10 min. This spatiotemporal rule has been used throughout transport studies developing stop & go algorithms for GSM data (Kim & Kwan, 2003; Bayir, et al., 2010; Iqbal, et al., 2014). Schlaich et al. (2010) proposed the use a 60 min rule *. For the previous example, using the 10 min rule, the user would have time to participate in an activity but using the rule of Schlaich, he/she would not.

* Mezuro released in April 2014 a preliminary dataset that used a stop & go algorithm in order to more accurately determine users’ activities. The result of this analysis is an Origin-Destination matrix and the method used is called ‘framing’. Mezuro used a 20 min spatiotemporal rule.

The core of the stop & go algorithms is the spatiotemporal rules. Another example comes from Rorije (2011) who listed numerous rules of thumb for identifying modal split. However, these require sophisticated route matching algorithms which are also a matter of research (Laasonen, 2005; Caceres, et al., 2007; Schlaich, et al., 2010). Determining mode choice, which is of extreme importance for creating OD matrices, is a main topic in current research (Reddy, et al., 2010; Wang, et al., 2010).

An example of a stop & go algorithm is the one used in the framing data. The ‘framing’ algorithm of Mezuro is used to determine when the user connects to two antennas whose service areas overlap. This point is saved as a potential place that the user stopped to participate in an activity. To determine if that is so, the next point of the user, where he/she connected to two antennas with overlapping service areas, is analysed. Two consecutive in time points determine the spatial and temporal length of a frame. After determining the frame, the algorithm determines if the user stopped at each start and end points of the frame.

The creation of the OD matrix from GSM data is not a complicated process once the trip’s origin and destination has been determined. This may be the reason why current research leans towards
establishing the methodology to determine origins and destinations as well as the modes of transport. In the case of this thesis, the methodology to determine trips is out of scope because it lies solely in the field of Mezuro B.V.

### 2.4 Other mobility studies using GSM data

Smoreda et al. (2013) noticed that the spatial patterns of movements, which they called motif and that are calculated from household surveys, are almost the same between Paris and Chicago. Additionally, the authors measured the observed patterns using individual GSM data for the city of Paris and found that they are also similar to those found by the household surveys. The motifs are network movements created by the trips of the users and are presented in Figure 3.

The patterns can be behavioural rules for travellers for use at agent based modelling. In strategic modelling only one trip is studied. GSM data show the possibility to extend this view and provide insight into the diversification of travellers. Nevertheless, it should be noted that the spatial resolution is of importance. Figure 3 shows motifs about travels within a city. It is unclear whether these patterns apply for higher resolutions such as municipalities. Smoreda et al. (2013) with their work provided a way to determine the accuracy of analysing GSM data; transportation studies using GSM data should expect to determine similar motifs.

Creating an algorithm that can effectively reproduce the origin destination patterns requires the determination of some locations such as the home location. The determination of the exact location of the home of the user is bound to accuracy problems. According to Keij (2013), the home and work location can be estimated using the method described by Maestre et al. (2009). Maestre et al. (2009) developed a “geospatial-temporal commuting model” to study commuting patterns. The authors were able to determine the movement of commuters using a simple geospatial model and identified the patterns with GIS. However, the authors do not determine the accuracy of their model because they do not relate their findings to socio-economic data.

![Motifs from household surveys and GSM data](image.png)

*Figure 6: Motifs from household surveys and GSM data, Source: Smoreda et al. (2013)*

Another algorithm for determining the residence of a user was proposed by Frias-Martinez et al. (2010). The authors used Genetic Algorithms to determine the residential locations. The algorithm determines the accuracy of locating the residence of a user by comparing the zip-code of the user’s home location and the antenna that the user is connected to. This is done using spatiotemporal rules. For example, when the user is found to make a call during a time interval, the algorithm determines the antennas that this user was connected to at the same time interval at different days. This creates alternative solutions for the time span of the dataset. The alternative solutions are
examined using a “fitness function” which takes into consideration the coverage and accuracy of each solution. Using an iterative process of comparing individuals of random samples simultaneously, the algorithm reaches convergence by not being able to find individuals that have a higher fitness value than those already examined. Finally, it should be noted, that coverage is actually the percentage of users that are assigned a cell tower as place of residence. This requires a coverage map which was created by the authors using Voronoi’s algorithm (Voronoi, 2009). The final results show that determining home location using GSM data is only 64.51% accurate when using 1 month data. Frias-Martinez et al. (2010) also used the same algorithm for data collected in a matter of more months and showed that the accuracy only marginally increases with bigger datasets. More specifically, using 4-month data, the accuracy increased to 67.21%. The authors also mention the computational time of this algorithm which is 1.15sec for evaluating one individual.

2.5 The understudied database of Dat.Mobility B.V.

This chapter will discuss briefly the databases that are of interest for this thesis. Additionally, the limitations of the database that were found based on the definition of the various parts and the general characteristics of the database are discussed and listed.

The two databases that will be discussed in this chapter have a major difference in the level of information’s aggregation. GSM data can be characterized by being in aggregate or individual level. Data are in individual form when the CDR contain information about an individual’s events and are in an aggregated form when they contain information about events in an area. A list of the most important categories of mobile phone data is shown in Table 1.

Table 1: Type of mobile phone location data, Source: Calabrese (2011)

<table>
<thead>
<tr>
<th>Type of mobile phone location data</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregated cell statistics</td>
<td>Easy to manage, Possibly in real time</td>
<td>No information on users’ mobility</td>
<td>Land use estimation, Population density estimation</td>
</tr>
<tr>
<td>Aggregated CDR with cell location information</td>
<td>Easy to manage</td>
<td>No individual interaction information</td>
<td>Connection between places, Regional partitioning</td>
</tr>
<tr>
<td>Individual CDR</td>
<td>Individual communication patterns</td>
<td>Large dataset, Mostly not real time</td>
<td>Social network analysis</td>
</tr>
<tr>
<td>Individual CDR with Cell location information</td>
<td>Individual communication and mobility patterns</td>
<td>Large dataset, Mostly not real time</td>
<td>Mobility analysis between large areas</td>
</tr>
<tr>
<td>Individual Event-driven data with triangulated location</td>
<td>Individual mobility patterns, Possibly in real time</td>
<td>Large dataset, Possibly need for special hardware to access data</td>
<td>Origin destination, Transportation mode</td>
</tr>
<tr>
<td>Individual Network-driven data</td>
<td>Individual mobility patterns, Possibly in real time</td>
<td>Large datasets, Possibly need for special hardware to access data</td>
<td>Useful for mobility analysis between large areas</td>
</tr>
</tbody>
</table>

2.5.1 The Mezuro Data

The data that this thesis will use are the outcome of the analysis of Mezuro B.V. Mezuro uses the raw data from a mobile network operator. These data are labelled Mezuro data. Mezuro data are event-driven data that can be categorized as “individual CDR with cell location information” (Keij,
and consist of 3 event types: voice call, SMS and data transfer, with information as to which user initiated or terminated a call. The events of Mezuro’s dataset are:

- Data event – No originating or terminating end
- SMS event – Originating (send) and terminating (receive) end
- Voice Call event – Place and receive a call / Terminate the call at either end

Therefore there are 7 events which can be used to determine the location of a user. A complete set of Mezuro data for 4 individuals can be seen in Table 2.

<table>
<thead>
<tr>
<th>Date &amp; Time</th>
<th>Hashed ID</th>
<th>Type of Call</th>
<th>Direction</th>
<th>Country Code</th>
<th>Cell ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-1-2013 14:37</td>
<td>5271410</td>
<td>S</td>
<td>MT</td>
<td>204</td>
<td>4227013</td>
</tr>
<tr>
<td>7-1-2013 15:25</td>
<td>7550920</td>
<td>V</td>
<td>TE</td>
<td>204</td>
<td>2028111</td>
</tr>
<tr>
<td>7-1-2013 12:06</td>
<td>7170230</td>
<td>D</td>
<td>U</td>
<td>204</td>
<td>12747674</td>
</tr>
<tr>
<td>7-1-2013 13:32</td>
<td>7327640</td>
<td>V</td>
<td>MO</td>
<td>204</td>
<td>5306942</td>
</tr>
</tbody>
</table>

It must be noted that, the phone number of each user is not shown but is replaced by a hashed ID. The hashed ID is replaced regularly so that a specific user cannot be traced. It is impossible to decrypt the hashed ID without information from the operator. The hashed IDs are updated regularly and are never the same for the same user. This guarantees that the data are always anonymized.

2.5.2 The viewDat database

The data that Dat.Mobility B.V. receives from Mezuro B.V. will be named for the remainder of this thesis as viewDat database or viewDat data. The best correspondence of the viewDat data to the categories of Table 1 is shown in Figure 7.

Because of the analysis that takes place by Mezuro B.V., the data of the viewDat database are heavily altered resulting in major information loss. The viewDat data are population counters showing how many people were located in a region. More specifically, they show the aggregated number of devices that created an event in the region. Because of the aggregation, it is impossible to follow a person’s movements. Dat.Mobility B.V. is acquiring the viewDat database every month since the 1st of January 2013.

The viewDat database is constituted of two datasets. The HPP (Herkomst Per Periode) and the DPU (Dichtheid Per Uur). The HPP dataset contains information about where the user created an event and where his/her home is located. Each sighting of the user is counted uniquely thus, the user cannot be counted more than once in the same region in the same time interval. The DPU dataset does not have any origin information. Additionally, the sightings of the users are unique per time interval/hour. Consequently, the user will only be counted once in the region where he/she
made the most events. Both DPU and HPP datasets are spatially aggregated in BAG regions * (Basisregistraties Addressen en Gebouwen).

By the start of this thesis, the viewDat database spans for 12 months, from January to December 2013. New data have become available in the process of this thesis. More detailed information about the viewDat database with a complete description of the characteristics and with samples of the datasets can be found in Appendix B.

* For more information about BAG regions, the available documentation can be found online at: http://www.kadaster.nl/web/Themas/themaartikel/Overzicht-BAGartikelen/BAG-documentatie.htm [In Dutch] and the geo coded system can be found online at: http://bagviewer.geodan.nl/index.html

2.5.3 A preliminary view on the limitations of the viewDat database

The data discussed above are massive in size; there are more than 3 million events logged per hour on average (Keij, 2013) created by 5.3 million subscribers of the mobile network operator.

An important limitation is that any data about residents or total visitors should be aggregated so that more than 15 users are counted. If a region generates less than 15 travellers, then the origin of these travellers is labelled as 'unknown'. Travellers from such regions are summed in a category called 'unknown origin'. Consequently, this limitation concerns the HPP dataset alone since these travellers do not appear at all in the DPU dataset. This problem is coined as the ‘Less than 15’ problem. The term will be used interchangeably with the term trips of ‘unknown origin’. The reason for its existence is so that a single user cannot be traced.

A preliminary view on the limitations of the viewDat database, are summarized in Table 3.

Table 3: Summary of limitations of viewDat database

<table>
<thead>
<tr>
<th></th>
<th>DPU data</th>
<th>HPP dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall</strong></td>
<td>Spatial aggregation is coarse as compared to other models (e.g., LMS)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>All counts must be above 15. If the counts are less, then there is a reading of 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>There are missing values from the dataset of year 2013</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No indication for activity or which visitor participated in an activity</td>
<td>No indication if a visitor was passing by or remained in the area</td>
</tr>
<tr>
<td></td>
<td>Not possible to sum events for a whole day</td>
<td>Origins are home locations/ There are no trip data</td>
</tr>
<tr>
<td><strong>Temporal aggregation</strong></td>
<td>There is not a 24hr period</td>
<td>Periods do not contain peak hours</td>
</tr>
<tr>
<td><strong>Type of visitor</strong></td>
<td>It is not clear if the division to categories is done based on visits on workdays, weekends or both</td>
<td>Trip purpose is not clear or not clearly derived</td>
</tr>
</tbody>
</table>

Another limitation not listed above is that the amount of users after nightfall and until early hours is less than the amount of users in morning to evening hours. This can logically be attributed to the fact that users are at their home location after work hours and are not making a lot of calls. However, from the data perspective, this low amount of calls means that the user cannot be
located. The initial assumption is that a device always corresponds to one and only user meaning that the user always keeps his phone in close proximity so that the phone becomes a proxy of the user's movements. Therefore, if the phone is not located it is assumed that the user is not moving.

One of the strong points of ViewDat database, in terms of data accuracy, should be determining home locations. This is stated because the temporal association rule that is used by Mezuro B.V. has been used in previous research and because determining the location of the residence should be more accurate in spatial terms since a user is seen multiple times in the same area.

### 2.6 Setting the Initial Assumptions

The advantage of mobile phone data is that the size of the dataset in users and movements is enormous. In essence, mobile phone data give the opportunity to eliminate travel diaries because they are themselves a sort of travel diaries. The accuracy of the data is always improved while its inaccuracy is maybe comparable to that of hand-written travel diaries.

However, the data have some limitations as any other mobile phone dataset. These need to be taken into consideration in order to formulate the initial assumption. First of all, the sample used is subject to sample bias. As discussed in paragraph 2.5.1, Mezuro uses a large mobile network operator’s data. This sample is eventually a biased sample because only the subscribers of the mobile network operator are included which may have a different profile than subscribers of other companies such as KPN or T-Mobile.

Another factor that contributes to sample bias is the fact that users who do not own a phone are excluded from the data. Keij (2013) states that the dominant event which Mezuro identifies in the mobile network operator’s dataset, is mobile internet data exchange. However, mobile internet data usage is not the same over the total population. Elder people’s (>65 years old) mobile internet use is little more than 20% while mobile internet usage of younger generations (12 to 25 years old) is almost 85%. Consequently, younger people should appear as more active and located more times that elder people (Centraal Bureau voor de Statistiek, 2012). User activity can be correlated with multiple socio-economic criteria such as age, income or sex which cannot be studied because of the inability to relate these criteria to mobile phone data due to privacy restrictions.

Finally, using mobile phone data to determine user mobility results in more active users to be seen as more mobile (Wang, et al., 2014). This problem is similar but not identical to the aforementioned limitation, which was that not all locations are captured for less active users. The penetration of mobile phones in the Netherlands is approximately 118% (Centraal Bureau voor de Statistiek, 2014). Developed countries usually have a penetration rate above 100% which means that some users have two (or more) mobile phones. The case where a user owns two active mobile phones may result in the problem that some users are over-represented or that a larger amount of trips is measured than are actually there (Wang, et al., 2014). This may also have other implications. For example, it is not uncommon that users own a second mobile phone that use for work related purposes. It is possible then that this phone’s movements describe work related trips more accurately and that the other phone does not represent work related trips at all.

Considering the above limitations, this thesis will make two assumptions:

- Sample bias can be looked over because the sample is very large.
- The mobile phone can be used as a probe for the user’s movements. This can be broken down to two effects:
  - The first effect is that the position of the mobile phone is the position of the user. This may not always be the case. Patel et al. (2006) showed from investigating 16 individuals over a period of 2 weeks that on average, users had their phone within “arm’s reach” 70% of the day time when away from home and 50% when at home. The percentage of the day time when the
phone and the user were in the same room was almost the same for both occasions (around 80%). Patel et al. (2006) also listed factors that affect the proximity of users to their mobile phones. Some of them are: routine (e.g., always place phone at kitchen counter when home), environment (e.g., in a car), physical activity, disruption to others (e.g., meetings), disruption to self (e.g., sleeping), regulations (e.g., in airplane), quick trips (e.g., leave the phone at desk during coffee breaks). More research on verifying this assumption is required.

- The second effect of the assumption is that one mobile phone is one user. Therefore if two mobile phones share the same movements, they will be considered as two users. Consequently, the terms device, user, individual and traveller are the same under the assumption and will be used interchangeably for the rest of the thesis.

## 2.7 Conclusions

Although research work on individual GSM data is extensive, the research on aggregated GSM data is scarce or refers to an obscure goal of studying population dynamics. However, researchers that used aggregated GSM data were able to demonstrate that GSM datasets show movements of people that are also observed in situ.

The most important point in current research is the derivation of Origin-Destination matrices which is clearly achievable using individual GSM data. Additionally, the privacy issue of deriving such information from GSM data is not factored in any study that was encountered. If the legislation or self-regulation of companies shifts to a more aggressive stance against “violators” of privacy issues, then the question would become what is the least amount of information required to derive origin-destination information.

This thesis will focus on a variant of this question which is where current research seems to lag behind. The dataset that will be used is much simpler than individual CDR used by most of research work on the topic but more advanced than Erlang data used by Ratti et al. (2006). In essence, there are two main differences from Erlang data and other population counters on an aggregated level:

1. The HPP dataset contains information about home locations and.
2. The DPU dataset contains information about where the user has been sighted the most.
3. Data analysis

3.1 Introductory Remarks

This chapter has three main goals. First goal is to check the database for consistency. Checking for consistency requires the examination of the monthly variation of the database for any of the groups of visitors or residents. The residents’ group should be more robustly calculated for reasons explained in 2.5.3. Therefore the first part will focus mostly on this group.

The second part of this chapter attempts to determine if one of the user categories of the database could be connected to a trip purpose. This would greatly assist in the creation of an OD matrix. The user categories in the viewDat database are created based on the number of times that a person visits an area. If he/she visits the area more than 10 times, he/she is considered a frequent visitor, if he/she visits 3-10 times, a regular visitor and if he/she visits less than or equal to 2 times, he/she is considered an incidental visitor. In this part of the chapter, the visitors’ group will be analysed. In doing so, the size of the sample should also be estimated because it will help in scaling the sample of phone users to the population of the Netherlands.

The last part of this chapter discusses in detail the limitations and issues of the database. This is important in order to create strong assumptions that will lead to the calculation of the OD matrix as it is done in chapter 4.

3.2 Analysing the total number of residents

The data provide the possibility to accurately determine the location of the home place of mobile phone users. The method that Mezuro B.V. uses is to recognize the location where a user is found most of times in a month in the time interval of 8:00 p.m. to 7:00 a.m. Due to the fact that the user is repeatedly found in the same location, the estimation of the home place should be more accurate than determining the amount of visitors. However, there have been indications that even with data for more than a month, estimating the home location may not be accurate (Frias-Martinez, et al., 2010).
In this subchapter the accurate location of the user is not important since the total amount of residents is used. Therefore the accuracy relies on the temporal rule used to determine if a person is a resident which in the case of the data from Mezuro B.V., is consistent with proposed temporal rules found in literature (Calabrese, et al., 2011; Berlingerio, et al., 2013; Fang, et al., 2014).

The average residents per month is shown in Table 4 and visualized in Figure 8.

**Table 4: Average number of residents per month**

<table>
<thead>
<tr>
<th>Month</th>
<th>Amount of residents</th>
<th>Month</th>
<th>Amount of residents</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>2,614,349</td>
<td>July</td>
<td>2,434,262</td>
</tr>
<tr>
<td>February</td>
<td>2,664,318</td>
<td>August</td>
<td>2,379,661</td>
</tr>
<tr>
<td>March</td>
<td>2,622,095</td>
<td>September</td>
<td>2,597,092</td>
</tr>
<tr>
<td>April</td>
<td>2,363,018</td>
<td>October</td>
<td>2,581,524</td>
</tr>
<tr>
<td>May</td>
<td>2,646,573</td>
<td>November</td>
<td>2,500,197</td>
</tr>
<tr>
<td>June</td>
<td>2,768,380</td>
<td>December</td>
<td>2,495,878</td>
</tr>
</tbody>
</table>

The amount of residents per month fluctuates by an average of approximately 100,000 residents. This is almost 4% of the average amount of residents. However, the amount of residents seems reduced in November and December but the reduction is not big enough to signify an inconsistency. On the other hand, the reduction from March to April and the increase from April to June seem peculiar and cannot be intuitively explained.

The amount of residents is reduced in the weekends. This should be attributed first to the fact that some people leave their homes on weekends for vacation and second to the phone usage which should generally be reduced in the weekends. Another interesting observation is that the number of residents is reduced in July and August. This was expected since these are the two months when a lot of people go on vacation either in the Netherlands, e.g., to a vacation house, or abroad. Therefore, they would stop being categorized as residents.

The information that were derived and that are of use later on, is the average number of people who are categorized as residents which is 2,555,612 residents. This number is much lower than the subscribers of the mobile network operator which are 5.3 million users (Keij, 2013).
Mezuro B.V. made the information on active users available in March 2014. The active users are devices that appeared in the dataset of the mobile network operator per day. The data are shown in Figure 9.

Figure 9: Average number of active users for all the Netherlands per month

Figure 9 is comparable to Figure 8 which was expected. However, there seems to be a problem with April and November. The amount of residents should have increased from October to November and shouldn’t have decreased from March to April. This could indicate that there is an inconsistency with Mezuro’s algorithm for these months.

Additionally, months before and including June seem to be the most active in terms of events. Therefore, these months should show an increased number of movements and should be used for mobility analysis. In any case, months of July and August will show a decreased number of movements probably attributed to summer vacations and should not be used from mobility studies.

The last part of the analysis of the residents is determining the outliers of the dataset and more importantly the frequency of appearance. Intuitively, there shouldn’t be a difference in the amount of residents in weekends since the vast majority of individuals should appear in their homes at that time. However, there is such a variation (see Figure 14, Figure 15 and Figure 16). It would be mathematically correct to exclude weekends from checking for outliers but it is interesting to see how many outliers appear in weekends. The outliers were determined using the Interquartile range-IQR which is the difference between the upper and lower quartiles. The results are shown in Table 5. Due to the seasonal variation the outliers were determined per month and not in the total of the dataset.

Table 5: Amount of outliers per month

<table>
<thead>
<tr>
<th>Month</th>
<th>Amount of outliers</th>
<th>Outliers in weekends</th>
<th>Month</th>
<th>Amount of outliers</th>
<th>Outliers in weekends</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>1</td>
<td>-</td>
<td>July</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>February</td>
<td>2</td>
<td>2</td>
<td>August</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>March</td>
<td>-</td>
<td>-</td>
<td>September</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>April</td>
<td>-</td>
<td>-</td>
<td>October</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>May</td>
<td>-</td>
<td>-</td>
<td>November</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>June</td>
<td>6</td>
<td>4</td>
<td>December</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
3.3 Analysing the number of residents per location

A question that arose from the data analysis in subchapter 3.2 is the difference of people that were categorized as residents in the GSM dataset to the actual residents of the regions. The population counts were drawn from the 4 NRM models whose zones can be aggregated to BAG regions. Figure 10 portrays the difference as a percentage for data aggregated per day type. Each population datum per day and region for the viewDat database is divided to the equivalent(s) from NRM and the difference per day type is averaged.

Figure 10: The difference of residents from GSM to actual residents of regions of the Netherlands

Figure 10 again validates that the estimation of residents in weekends is reduced. This strongly suggests that the phone activity plays a vital role in estimating the residents. The reason for the reduced number of residents on Mondays is that the datum points of GSM data of regions which overestimated the number of residents were excluded. There were a total of 4,241 datum points (out of 64,455 data entries) that overestimated the number of residents. There is no pattern for this overestimation meaning that it is not focused on specific regions or months but it always occurs for regions that have a low number of residents (below 1,000 registered residents).

The spatial variation of residents should better be shown by Figure 11. In this case, the residents per region were averaged per month and were compared to the residents from NRM data. Only data for July are shown to reduce graph complexity. The formula that was used to determine each value of each axis is the following:

\[ y,x = \left( \frac{Res_i}{Res_{tot}} \right)_{NRM,GSM} \]  

where, \( y,x \) are the values for axis y and x representing NRM and GSM data respectively and \( Res \) is the amount of residents for regions \( i \).

The formula is calculated per region. It was noticed that some regions had two codes. Because it could not be identified which one is correct, both entries for these regions were deleted. The total population from NRM was 15,412,537 which is a reduction of approximately a million individuals from the total population.
The variables shown in Figure 11 are clearly related. A line was fit to the data using simple linear regression and the fit was tested using $R^2$ which can be found on the graph. This graph shows that the percentage of a population of a region over the total population of the dataset is approximately equal for the same regions when calculated for both datasets. For example, if 5% of the total population of the Netherlands resides in a region, then the amount of residents found in the GSM dataset is 5% of the total sample of GSM data.

This can be argued because the linear regression line crosses or is near to points in the bisector of the Cartesian plane. That is an important argument against the bias of the dataset. If there was a systematic bias, one would expect that the regression line would diverge significantly from the bisector. Thus, the GSM data would consistently show less or more residents than are actually there. Especially, if the data were not consistent some regions would show more residents than they should and others less. Finally, if the bias was not systematic then the data would appear scattered. None of the two is the case which implies that the data are consistent.

The outliers that appear in Figure 11 show how the spatial variation of the mobile network operator’s market share, affects the data. The 3 outliers at the top of the graph show Amsterdam, Rotterdam and The Hague respectively. Since the residents of Amsterdam show a higher percentage from the GSM dataset than from the NRM dataset, then there should be more users of the mobile network operator in Amsterdam and vice versa for Rotterdam and The Hague.

Next step is to attempt an estimation that would raise the number of devices to people. Because of Figure 11, there is a good indication that this relationship will be linear. Simple linear regression analysis was run for the dataset from July to December 2013 for each day type. The analysis maybe should be done in a less temporally aggregated setup but even with the current setup, the results appear to have significant statistical power as shown in Table 6.

The fit of the linear regression lines to the data is very good as determined by $R^2$. The equations can be used per day type to estimate the number of people given the number of devices.
### Table 6: Coefficients of Linear Regression

<table>
<thead>
<tr>
<th>Day Type</th>
<th>Coefficient $\alpha$</th>
<th>Intercept $\beta$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday</td>
<td>6.046</td>
<td>1719.86</td>
<td>0.8</td>
</tr>
<tr>
<td>Monday</td>
<td>5.197</td>
<td>1797.85</td>
<td>0.8</td>
</tr>
<tr>
<td>Tuesday</td>
<td>5.188</td>
<td>1819.67</td>
<td>0.8</td>
</tr>
<tr>
<td>Wednesday</td>
<td>5.19</td>
<td>1794.98</td>
<td>0.8</td>
</tr>
<tr>
<td>Thursday</td>
<td>5.224</td>
<td>1798.6</td>
<td>0.8</td>
</tr>
<tr>
<td>Friday</td>
<td>5.279</td>
<td>1767.71</td>
<td>0.79</td>
</tr>
<tr>
<td>Saturday</td>
<td>5.914</td>
<td>1422.583</td>
<td>0.8</td>
</tr>
</tbody>
</table>

### 3.4 Analysing the visitors’ group of the viewDat database

The second larger group in the viewDat database is the visitors. A visitor is a person/device that made an event outside the area which is considered his/her home. Essentially, each visitor is a traveller because he/she has definitely travelled to reach another area.

Looking into the visitors’ group of data may give insight into the trip purpose. This is because each visitor is categorized based on the times he/she visited an area (frequent – $>10$ times / regular – $3$-$10$ times / incidental – $1$-$2$ times). Therefore, it could be that frequent visitors are in fact persons travelling with the purpose of work. This chapter will determine whether assumptions, such as the one mentioned before, are valid.

#### 3.4.1 Examining the frequent visitors of the 0-24h period

A separate experiment on the 0-24h period data attempted to predict the following month by applying time series analysis on each month. More specifically, by plotting the data for frequent visitors it was observed that there is systematic variation within a month that had a clear trend. Data were stable at a great extent within a week followed by a drop in weekends creating a seasonal pattern. This repeating pattern was found in almost every week of every month speaking for the consistency of the data.

Due to this variation a time series model could be applied. The purpose was to predict the following month each time which would lay the ground for multiple statistical analysis in the future. The 95% confidence interval of the prediction was also drawn. In order to have trust about the prediction, the data of the next month would have to lie within the 95% confidence limit. The time series model that was used is the simple multiplicative model which has the following generic formula:

$$ Y_t = T_t \times S_t \times I_t \times \varepsilon_t $$

where, $Y_t$ is the time series model, $S_t$ is the seasonal component, $I_t$ is the irregular component and $\varepsilon_t$ is the residual between the time series and the data.

Figure 12 shows that the time series model could not predict August using input from July. What is more interesting from the analysis is that it revealed an unsystematic or irregular variation within a year.

For that purpose, the trend component was separated and redrawn for the whole year of 2013. This was plotted in Figure 13 along with the actual data variation of visitors of all months. The corrected December refers to the time series model of only the first three weeks of December. The last week showed an important reduction which is attributed to Christmas vacations.
Chapter 3 - Data analysis

3.4.2 Comparison of the Trip Purpose graphs with viewDat data variations

Figure 14, Figure 15 and Figure 16 show the weekly variation of the 3 different categories of visitors that appear in the viewDat database. The figures are drawn for the average of all regions in the dataset. It is observed that the weekly variation of frequent and regular visitors have a stable value for working days and a reduced value for weekends. Especially, Sunday consistently shows the lowest amount of visitors. The picture is reversed for incidental visitors where the highest value is for Saturday and does not have a plateau for either working days or weekends.
The figures are comparable to the trip distributions per purpose per week day. The data are drawn from the Monitoringtool 2009. The Monitoringtool distinguishes 8 travel purposes and a category named ‘Other’. The two graphs in Figure 17 show which trip purposes demonstrate a similar pattern as seen in the GSM dataset.

Figure 14: Weekly variation of the frequent visitors for September and October

Figure 15: Weekly variation of the regular visitors for September and October

Figure 16: Weekly variation of the incidental visitors for September and October
In general, as expected, all recreational purposes show a pick of the number of trips in the weekends and a reduced amount in working days. Incidental visitors’ category shows the exact same pattern. This might be surprising because the amount of recreational and especially shopping trips is high. Since a person does more trips he should visit some areas more often therefore, qualify for regular or frequent visitor. Two explanations can be given: (i) the trip distance, especially for shopping, is short, as shown in Figure 18 which consequently means that person doesn’t travel far enough to be registered in another area, (ii) a person doesn’t use his phone often when in a recreational activity. The second explanation seems more plausible. If it holds, it would mean that it is impossible to determine recreational or shopping purpose in the trip patterns. This is opposite to some targets set by Mezuro which concern the analysis of shopping activities (see Appendix B).

Work and education related trips create almost a plateau for the working days, the same as in the frequent and regular visitors’ category, while there is a clear reduction of trips in the weekends. The general comparison of the figures above insinuates that there is a relation between purposes and type of visitor which is shown in Table 7.

However, in the absence of socio-economic data, to safely conclude that the relation of trip purpose and type of visitor actually holds is impossible. The comparison in Table 7 gives only an indication of the relation and is definitely not conclusive. The general assumption that, for example, frequent visitors are workers or incidental visitors are shoppers is a fallacy known as ecological fallacy (or fallacy of division in inferential logic) which occurs when the individuals’ nature is deduced from the
aggregate group that these individuals are believed to belong to. This problem is also mentioned by Calabrese et al. (2013).

Ecological fallacy has multiple facets one of which is Simpson’s paradox (Simpson, 1951) that might be at play in this case. Cohen and Nagel (1934) noticed that combining the rates or trends of two groups of data may result in a completely different and even reverse trend. The existence of Simpson’s paradox makes inference of the characteristics of individuals impossible to determine because the causal relationship has disappeared due to the combination of the groups.

Nevertheless, it is important to note that there seems to be a temporal relation between the frequency of visits and the purpose. More research is required to establish such a relationship although its foundation was given here.

Table 7: Possible relation of trip purpose and type of visitor based only on the shape of the variation

<table>
<thead>
<tr>
<th></th>
<th>Frequent visitors</th>
<th>Regular Visitors</th>
<th>Incidental Visitors</th>
</tr>
</thead>
<tbody>
<tr>
<td>To and from work</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Business trip</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Personal Trip</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Shopping</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Education</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Visiting/ Stay</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Social/ Recreation</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Touring/ Hiking</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Other</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
</tbody>
</table>

Figure 18: Trip length distribution and Trip distribution per purpose
3.5 Limitations and issues of the dataset

This chapter concludes with the issues and limitations of the viewDat dataset as they were seen so far. Discussing these limitations will be of great importance to efficiently estimate an OD matrix and to comprehend where potential improvements should focus that can improve the accuracy of the dataset. These issues are summarized in Table 8 and discussed in detail below.

Table 8: Summary of issues with the databases that damage the quality of the OD matrix

<table>
<thead>
<tr>
<th>Database</th>
<th>Issue</th>
<th>Description</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPP</td>
<td>Specification of origin</td>
<td>The origin is always home location</td>
<td>(a)</td>
</tr>
<tr>
<td></td>
<td>‘Excess’ trips</td>
<td>Because of using period instead of an hour, there should be extra trips in the database besides work/education related trips</td>
<td>(b)</td>
</tr>
<tr>
<td></td>
<td>Multiple counts</td>
<td>There are multiple counts of the same user in a period</td>
<td>(c)</td>
</tr>
<tr>
<td>DPU</td>
<td>Not always end of trip</td>
<td>Sightings are not only people that reached their destination but may include others still travelling</td>
<td>(d)</td>
</tr>
<tr>
<td>Both</td>
<td>Spatial Aggregation</td>
<td>The spatial aggregation is not adequate for cities and is too detailed for other areas that do not generate or attract a substantial number of trips</td>
<td>(e)</td>
</tr>
<tr>
<td>Datasets</td>
<td>‘Less than 15’</td>
<td>The ‘Less than 15’ problem damages the quality of the matrix for determining trips in small or scarcely populated regions</td>
<td>(f)</td>
</tr>
</tbody>
</table>

(a) Specification of origin

Because origin in the HPP dataset is always the home location, there is no possible method to directly extract other movements of users besides from home. Typically, hours besides morning peak, contain other trip purposes such as shopping and business trips whose origin can be work or another location (De Dios Ortúzar & Willumsen, 2011). Therefore, only a morning peak matrix can be created from the viewDat database.

(b) ‘Excess’ Trips

It is known that the HPP dataset is divided in periods and that the sightings of users are not unique. Therefore, each time period will include multiple trips (named here ‘excess’ trips) of the same person possibly for different purposes. A long time period will include more ‘excess’ trips than a short time period. Nevertheless, the time of day that the time period spans to, also plays a role to how many ‘excess’ trips are found in each period. For example, the morning peak (7-9h) includes predominantly home-to-work trips (De Dios Ortúzar & Willumsen, 2011). Period 10-12h has a completely different profile with trips for multiple purposes which are usually shorter in time and distance. Thus, the time period 10-12h may include multiple trips for different purposes for the same user. This issue has exactly the same effects on the main research topic, which is the derivation of OD matrices, as described below for multiple counts but while the ‘excess’ trips are attributed to the temporal inconsistency of the DPU and HPP datasets and to the definition of the HPP dataset, the multiple counts are attributed only to the latter.
(c) Multiple counts

By definition the HPP dataset will have multiple counts of the same user within the examined time interval. When a user creates an event, he/she is logged uniquely per region. Thus, if the user crosses multiple regions at the examined interval and creates an event in each, he/she will be logged in all of them. Estimating the extent of this issue can find uses in determining methods to scale the sample of the HPP dataset to the amount of users of the mobile network operator.

Estimating the number of multiple counts could be done with the mark-recapture technique which is also known as tag-release (Krebs, 2013). The goal of this method is to estimate the population of a species in an area by using a sample of the population. This can be achieved by capturing a sample of the population and tagging the animals for example, with collars. After a short amount of time, one would recapture another sample and measure how many tagged animals of the initial sample are recaptured. The population can be calculated by the following formula:

$$\frac{R}{T} = \frac{M}{N}$$  \hspace{1cm} (3)

where, \(R\) is the amount of tagged recaptured animals, \(T\) is the size of the recaptured sample, \(M\) is the amount of animals that were tagged in the initial sample and \(N\) is the size of the population.

Equation (3) is known as the Lincolnson-Petersen method (Seber, 1982). Whether this method fits in the problem at hand, is based on whether the assumptions of the method can be satisfied which are the following:

- Capture of animals should be random so that each individual has an equal chance of being in a sample. This assumption cannot be verified without knowing the phone activity and its dynamics. It is assumed that the phone activity is a random variable for the persons that are travelling.
- Nothing has occurred that has disturbed the size of the population. For example, no animal has died or no animal was born. For the case of GSM data, no new phone appeared or no user halted using his/her phone. This assumption can be satisfied by using the amount of residents as the size of the sample instead of the active users. Residents are people that have been identified multiple times within a month thus; the probability of them stopping their contract with The mobile network operator is minimal.
- The mark-recapture method assumes that the sample that was captured initially had the appropriate time to be dispersed in the population. This is a priori the case with GSM data since no actual capturing took place.
- The tag of the animal is not lost or destroyed. This also is satisfied with mobile phone data.

Applying equation (3) for two consecutive hours can provide the amount of people that appeared in both samples. In other words, it is assumed that the sightings in the DPU dataset (which are unique individuals) of an hour represent the captured sample. The sightings of the DPU dataset of the next hour represent the recapture. The two consecutive hours that will be examined are 10:00 a.m. (10-11h) and 11:00 a.m. (11-12h). From equation (3), \(R\) is the requested variable, \(T\) is the size of the sample in the second hour and equal to 1,306,694 individuals/devices, \(M\) is the sample of the first hour which is equal to 1,268,542 individuals/devices and \(N\) is the amount of travellers in the dataset which is 2,777,090 individuals/devices (in paragraph 3.6.1 it is argued that the size of the sample is equal to the amount of residents because they are the persons who travelled or had the potential to travel). Thus:

$$\frac{R}{T} = \frac{M}{N} \iff \frac{R}{DPU_{10-12h}} = \frac{DPU_{10-11h}}{HPP_{10-12h}} \iff \frac{R}{1,306,694} = \frac{1,268,542}{2,777,090} \Rightarrow R = 596,882$$

Therefore, 596,882 individuals created an event in two consecutive hours. Each of the two hours is comprised of the individuals who appeared only in one and those that appear in both hours. The total of unique individuals that appeared in the period 10-12h should be equal to individuals that
created an event only at 10h, those that created an event only at 11h and those that created an event at both hours which is equal to $R$. The former two numbers can be found by subtracting $R$ from the sightings of the DPU dataset for each hour respectively:

$$R + (DPU_{10-11h} - R) + (DPU_{11-12h} - R) = 596,882 + (1,268,542 - 596,882) + (1,306,694 - 596,882) = 1,978,354$$

The difference of the unique individuals to the HPP sightings is an indication of the amount of multiple counts. In other words, 21.5% of the sightings of the HPP database for Monday 3rd of March 2014 and for 10-12h can be accounted to multiple counting.

The effect of the multiple counts’ issue (and of the ‘excess’ trips) is threefold:

**Problem 1**

The following map shows the antennas that service the area of TU Delft. A mobile phone in TU Delft can connect to any of those antennas and consequently can be into any of the more than 20 BAG regions that these antennas are a part of.

![Antennas map](image)

*Figure 19: Antennas that a mobile device at TU Delft can connect to, Source: Keij (2013)*

Consequently, the amount of sightings is not the amount of users who created an event. It’s more problematic if one considers that a user might change antennas irregularly because of the network usage, i.e., without moving from his/her location. Thus, the user will be logged in a nearby area without travelling to that area. This problem should be manifested more regularly in larger time intervals.

This problem cannot be examined from the side of viewDat database but there are solutions that can be employed from the side of Mezuro as proposed by Becker et al. (2011) and Wasif (2012). In any case, it is essential to estimate the accuracy of the positioning algorithm in order to improve the movement patterns. Employing an algorithm that determines the activities of individuals would definitely eliminate this problem. However, in paragraph 4.4.1 it was found that trips below 7.5km appear as intra-zonal. Thus, the implications of this problem might not have any practical effects.
Problem 2
The multiple counts issue extends to the user categories (frequent, regular and incidental) of the viewDat database. For example, consider a user moving from Deventer to Utrecht. Both train and car modes cross approximately the same regions for this trip as shown in Figure 20. Since the user is making this trip almost every working day, it is possible that he/she will appear as a frequent visitor to both Amersfoort and/or Apeldoorn as long as any other region/city in between. Thus, the multiple counts issue is a strong argument that the user categories do not correspond to a specific purpose.

Problem 3
Because of multiple counts, the sightings per region are increased. But, the HPP dataset also contains information about the origins of persons per area. If a user is counted multiple times, then the origin of this user is also logged multiple times. To better illustrate this issue, consider the following example, 2 users begin a trip from an origin moving towards destinations/regions 1 and 2 respectively while region 2 is reachable only though region 1. These are shown in Figure 21.

All movements are analysed for the same time interval. Note that destinations and origins are only named as such. In the HPP dataset, they appear as counts. Additionally, the figures in the right show the counts of origins and destinations as they should appear by definition (classical modelling).
Both users create an event regularly enough to be seen in both regions. In both cases shown above, users have not started moving and thus the origin and destination count is 0.

In Figure 22, users have reached Region 1. One of the two users will continue moving to Region 2 while the other will not. In this case both users are counted in Destination 1 and their origin is logged. The first difference that appears is that the destination count (or the sighting values of the HPP dataset) is higher for the HPP dataset than what it should have been (right picture).

Both users have reached their final destination in Figure 23. Since the sightings in the HPP dataset are unique per region, the user that remained in Destination 1 does not increase the total sightings or origin counts. However, the user that reached Destination 2, increases both destination and origin counts.

**(d) Not always end of trip**

This issue is similar to the ‘excess’ trips of HPP dataset. It is possible that sightings of the DPU dataset refer to people that are still travelling.

**(e) The Spatial Aggregation**

The matrix that will be derived concerns the total of the Netherlands. The best comparison of the derived OD matrix can be with a model that also concerns the whole of the Netherlands. Such a matrix is the LMS matrix. However, there are limitations to this comparison. One important limitation
is that the spatial resolution of the LMS matrix does not correspond to BAG regions. There are
major differences in the way the zones are divided among those two systems in locations with low
population, e.g., Terschelling, and locations with high population, e.g., Amsterdam.

![Figure 24: Differences between BAG and LMS zoning system specification for Terschelling](image)

*Only the centroids of mentioned regions are included in the red circles*

![Figure 25: Differences between BAG and LMS zoning system specification for Amsterdam and Almere](image)

Figure 24 and Figure 25 show two important things. First, BAG regions’ specification is not just a
coarser spatial aggregation and second, they are not created, and possibly cannot be used, for
transportation studies. The reasoning for the latter argument is that a TAZ, such as the LMS zones,
are created in such a way that intra-zonal trips have no geographical reference (see subchapter
2.2). It cannot be argued that trips within the whole city of Amsterdam have no geographical
reference.

The reason that Mezuro used such large regions should be because the estimation of individuals
within cities is not accurate enough. These cities are serviced by numerous antennas whose
service areas overlap while some antenna’s area spans to more than 30km (Keij, 2013).
Additionally, outside the densely populated areas, the BAG regions’ specification is much more
detailed than it should which results in many regions not having any sightings because of the ‘Less
than 15’ problem.

(f) ‘Less than 15’ problem

The ‘Less than 15’ problem has been discussed in 2.5.3 and a solution is proposed in 4.3.2
3.6 Conclusions

A conclusion of this chapter is that the data of the viewDat database show seasonal variation while statistical values taken from external databases, e.g., CBS, usually do not; they have a flat value for a whole year. Therefore, comparing a variable that changes in time to a single value that remains stable within a year is not advised because conclusions about the underlying cause of the variation of a variable or of the definition of the variable itself are impossible to make. The consistency of the data from one month to the next is a better indication of the accuracy of the data.

Another problem can result from wrong inferences such as for trip purpose. Assuming that a category of visitors contains only a specific trip purpose, will result in ecological fallacy. An outcome of this fallacy is that, for example, if frequent visitors are assumed to be workers, then they will be assumed to vary temporally as the workers’ category. But this may not be the case and if it is, there will be no causal relationship.

It is also important to know the exact amount of the size of the sample. This can help in raising the sample to the population among others. Since it cannot be calculated from the side of Dat.Mobility, it should be calculated from the side of Mezuro. Furthermore, the amount of active users and of residents corresponds neither in absolute values nor in temporal variation. Mezuro has to determine why this takes place and if it is attributed to the algorithm for determining the residents.

Another conclusion is that the multiple counts problem damages the HPP dataset greatly. It should be noted that other purposes might be in play besides work but it is still a very high amount to be only attributed to other purposes. This problem relates to the definition of the HPP dataset which is that each time interval contains unique appearances of each user per region. That means that if a person works in Delft and travels for a business or other meeting in the Hague and then return to work in Delft, then he/she will only be counted only once in Delft for the 0-24h period at the HPP dataset. This is because the person has to uniquely appear in each location. This definition however creates more problems than it solves. The HPP dataset is actually shows the regions that a user visited and where his/her home location is. Thus, it is neither a counter of visitors nor it contains origin-destination information.

Finally, it was concluded from the data analysis that months before June and not including May should be excluded from further enquiry for the following reasons:

- Data for January and February 2013 show a reduced average amount of frequent visitors that cannot be explained (see Figure 13). Since these were the first months that the data were made available, the algorithm that Mezuro B.V. used should have been in its initial stages.
- Data for March show an abnormally increased number of frequent visitors (0-24h period) compared to the rest of the months (see Figure 13).
- Data for June appeared to have 6 outliers (see Table 5). In other words, 6 out of 30 days showed a value higher or lower than expected.
- Resident data for April plummet compared to the months before and after. The drop from March to April and the increase from April to May is not equivalent to the graph of active users but the amount of residents of May and March are on the same level as expected (see Figure 8 and Figure 9).
Resident data for November appeared to decrease from October when the active users are increase between these two months (see Figure 8 and Figure 9). It is not clear whether this is an inconsistency but there also does not seem to be a causal effect that would result in a reduced number of residents.

### 3.6.1 The problem of the sample’s size

The size of the sample is the amount of the mobile network operator’s subscribers that use their phones and create events in the spatiotemporal plain. Despite the available information, the size of the sample is actually unknown. There are 5 different sources from where one can draw an estimation of the size of the sample:

- **Keij conducted research on location based applications using the Mezuro dataset.** He notes that the size of the Mobile network operator’s sample, which is the basis for the Mezuro dataset, contains information about 5.7 mln users/subscribers. Nevertheless, when he calculates the ‘events per user’ factors, he estimates that in the interval 15-18p.m., which is when users appeared more active, a user creates 1.94 events (with standard deviation of 0.27). In that time frame, the total amount of events never surpassed 7 mln in an hour. Thus, he used a smaller amount of users than he initially stated to calculate the aforementioned values. *(Keij, 2013)*

- **The mobile network operator published a document about the financial results of the year 2013 in which it claims that the total subscribers in the Netherlands are 5.288 mln of whom, 32.2% or 1.7 mln use prepaid.**

- **Deloitte conducted a mobile consumer survey for the Netherlands *(Deloitte Holding B.V., 2013)* and claims that The mobile network operator’s share of the market in 2013 is 24%. The penetration of mobile phones in 2013 is 118%. Therefore the size of the sample according to Deloitte is 4,732,080 users.

- **On April 2014, Mezuro released the amount of ‘active users’.** Active users are the users that made an event during a day. The problem with the active users and estimating the sample size is that, it is not possible to know with certainty whether all users made an event during a number of days. Therefore it could be that users who appear as active in one day are different from active users of another day. The maximum number of active users of a month or a year should be a better approximation of the size of sample but even in that case, there is no guarantee that all users have appeared in the dataset. The average active users are 3,066,000 and the maximum is 4,415,265 users observed on Tuesday 31st of December 2013 (the maximum for all the dataset without excluding the months mentioned in 3.6 is 4,557,448 observed on Monday 14th of January 2013).

- **The amount of residents can also be used for the size of the sample.** The residents’ category concerns users that live in the Netherlands. As stated in subchapter 3.2, the amount of active users might include information about users that do not live in the Netherlands but own a Dutch mobile subscription. Users that are not categorized as residents somewhere in the Netherlands will, by definition, not make home-bound trips. The accuracy of the estimation of the residents however is bound to the accuracy of the methodology used for determining their location. In any case, this should be a good approximation of the size of the sample for the data that were used in this chapter. The amount of residents is also bound to the same problem as active users which is, people not appearing in the dataset every day but only in a subset of the days of a month. The average number of residents for all months is 2,555,612 while the maximum number of residents is 2,851,225 users and was observed on Wednesday 11th of December 2013 (the maximum for all the dataset without excluding the months mentioned in 3.6 is 3,634,580 users observed on Monday 14th of January 2014).

The best approximation for the sample size should be given by the active users’ amount because these are the people for whom the dataset is created. However, not all active users are included in the HPP and the DPU datasets of viewDat database. As mentioned above, if a person is not
registered as a resident, then he/she will not generate any trips whatsoever. This happens because all the counts of the ViewDat database contain information about trips from home. Therefore, if a person is not registered as a resident, he will never be counted in any location as a traveller. Consequently, the sample of people who participated in “trips” in the HPP dataset is the residents. This amount is named ‘lower limit’.

It was found in subchapters 3.2 and 3.4 that the amount of residents and visitors show systematic weekly variation and unsystematic/irregular yearly variation. The amount of residents and visitors drops in some months, for example July and August, and increases in others, for example September. This can be attributed to holidays. The effect of holidays is clearly obvious from Figure 8 and Figure 13; the drop in the last week of December, when the Christmas holidays take place, caused the average of December to drop significantly. In holidays, people may become untraceable as residents. The amount of residents discussed above concerns the amount of traceable users. Thus, this amount varies per day and month while it is always lower than the amount of traceable and untraceable users put together. The amount of all users traceable and not, is the amount of active users. To avoid the problem of devices not appearing every day because they didn’t create an event, the maximum amount of active users will be used which is 4,415,265 or approximately 24% of the mobile phone market. Note that this number is almost equal to Deloitte’s estimation discussed above (Deloitte Holding B.V., 2013). This number is called ‘upper limit’.

Concluding, there is not a consistent, throughout the days and months, amount to be used for the size of sample. It varies based on different circumstances mostly relating to phone usage and temporal variations. In between the lower and upper limit, there is degree of uncertainty because of the user simply not making a call or because a user is not in the Netherlands. Therefore, it is decided that the average amount of residents and the maximum amount of active users be used as limits.

If a single value should be selected to be used for the size of the sample, then this is not proposed to be the average of the two aforementioned limits because it would have no physical meaning. Instead the maximum number of residents should be used: 2,851,225 users. This amount presumably has the lowest amount of untraceable users while it concerns people that participated in trips. In other words, this is the largest amount of people that were located (not only appeared) in the dataset and participated or had the potential to participate in trips.
4. Acquiring Origin-Destination information from the viewDat database

4.1 Initial Remarks

Based on the issues that were determined in subchapter 3.5, the following implications have to be considered in order to create an OD matrix from the viewDat database:

- Because the origin is always home, only the morning peak matrix can be extracted. All other time periods of a day have different origins such as work.
- The multiple counts and 'excess' trips issue may be the biggest drawback in creating an OD matrix. The biggest issue might lie in the ratio of origins between regions (see example of problem 3, issue (c) in subchapter 3.5).
- Although there is a clear (home-bound) origin, the destination is not clear.

Since these implications cannot be addressed directly, a clear list of assumptions has to be created that will assist in overcoming the issues. In order to test for the accuracy of the derived OD matrix, it will be compared to observed and modelled values that are used in practice. The ultimate goal is to establish the accuracy of the assumptions. The OD will concern only one day of one month of data and specifically Monday 3rd of March 2014. The methodology to create an OD matrix from the viewDat database is provided in the form of a pseudo-algorithm in order to be easily recreated for other days.

Another goal of the process described in this chapter is to find how the accuracy of the OD matrix can be improved. Thus, this chapter will close with some recommendations for Mezuro.

4.2 List of Assumptions

Simply put, creating an OD matrix requires the derivation of origins and destinations. These cannot be clearly derived from the existing datasets. The previous subchapter showed that the HPP dataset contains counts and not origins nor destinations.

However, the information for extracting origins and destination is included into the datasets. Meaning, it is possible to derive this information by setting clear and logical assumptions.
accuracy of the derived OD matrix will be measured by the accuracy of the assumptions and how well the matrix performs when compared to observed values, such as the Monitoringstool 2009, and to other OD matrices, i.e., the LMS.

The following assumptions will be made in order to create an OD matrix. The list also includes the method that will be used to verify each assumption.

1. **The DPU sightings per region are the total destinations per region of the OD matrix**

   The destinations are retrieved from the DPU dataset. The DPU dataset shows actual sightings of users per hour that are unique; no user is double counted. Additionally, the users are counted in the region that they created the most events in an hour which, although it does not substitute the definition of activity, it can be used as a proxy for it.

   To strengthen the assumption, only the sightings of 10:00 a.m. will be used. In an aggregated sense, the amount of users seen in the DPU dataset after the morning peak period should concern travellers who reached their destination and travel either during or some time near the morning peak period. According to Ortúzar & Willumsen (2011, p. 141), the morning peak’s trip purposes are in majority (above 85% of total trips) related to work and education purposes. Some models distinguish work related and other purpose related trips to create a peak model (Crevo & Virkud, 1994) which shows how dominant the work related trips are in the morning peak hour model. Finally, using the hour immediately after the end of the peak period also relates to another phenomenon observed which is called peak spreading or peak broadening (Goodman, 1972; Hounsell, 1991). Multiple transportation models, including the LMS, use ‘peak-hour shoulders’ that extend to 10:00 a.m. with departure time choice (Chuck, 1999; Smit & Flikkema, 2010).

   **Verification:** The assumption will be verified by comparing the derived matrix to a morning peak matrix through the Trip Length Distribution – TLD. Additionally, specific trip rates per region will be compared to trip rates of other matrices.

   **Limitations:** Assumptions about work and education related trips being dominant in morning peak period, are used to generate trips, not to categorize them. It is easier to use a trip purpose in order to generate and distribute the trips in a manner that is coherent to reality. In the case of the viewDat database, the amount that will be used as destinations is an observed value which cannot be broken to purposes due to lack of, mostly socioeconomic, information. Thus the values will pragmatically include other purposes.

   Trips are usually calculated when modelling per trip purpose because different trip purposes can show different trip characteristics. Additionally, it has been found that trip purpose can affect mode choice (Racca & Ratledge, 2004; Driscoll, et al., 2013) and even route choice mostly though for walking and bicycling (Zhang & Levinson, 2008; Yang & Mesbah, 2013). These effects cannot be analysed by a matrix created from the viewDat database.

2. **The total origins per region are drawn from the HPP dataset for the 10:00a.m.-12:00p.m. period.**

   A smaller time period will have less ‘excess’ trips. Logically, it should also have a smaller amount of multiple counts. The time interval 10-12h is the smallest available time interval in the HPP dataset and is also in close time proximity to 10:00a.m. The 10-12h period concerns persons who have travelled within the morning peak and its shoulders and should have reached their destination.

   According to this assumption, the total origins per period of the HPP dataset will be used as origins of the newly formed OD matrix. There is a difference between the total sightings of the DPU dataset and the total origins of the HPP dataset (the latter is larger). Because of the high degree of uncertainty towards what the origins of the HPP dataset include, they will be equilibrated to the destinations (DPU sightings). This should reduce the amount of the ‘excess’ trips and multiple counts in the origins.

   **Verification:** Same as Assumption 1.
Limitations: Subchapter 3.5 showed how the multiple counts (and, using the same reasoning, the ‘excess’ trips), damage the information included in the origins of the HPP dataset. The problem is a matter of the amount of trips but most importantly of the spatial interactions in the, to be derived, OD matrix. Regions with more residents than workplaces are bound to generate more trips while regions with more workplaces will attract more. The former will therefore show an increased amount of origins due to the multiple counts than the latter. Consequently the ratio of generated trips of the two regions will not be as it should have been if the effects of multiple counts and ‘excess’ trips are alleviated. This problem cannot be solved without the use of a different database for origins or the improvement of the HPP dataset.

3. The origin-destination combinations of the HPP dataset contain accurate enough spatial information to be used as the OD pairs’ values.

Information about OD combinations in the HPP dataset should be as accurate as using the total origins per regions for the origins of the OD matrix (Assumption 2). There are two types of information that an OD matrix can give. First, how many trips took place between an origin and a destination and second, which regions had no trips between each other. The former is the same as in Assumption 2. The latter concerns the sparseness of the derived OD matrix.

Verification: Same as Assumption 1 and 2. Additionally, the sparseness of the matrix will be compared to the sparseness of other OD matrices.

Limitations: Same as Assumption 2.

Although it is not a formal assumption, it should be mentioned that the accuracy of the destinations of the OD matrix, which are the sightings of the DPU dataset, are considered more accurate than the information of the HPP dataset. The sightings of the DPU dataset refer to people while the HPP dataset contain information about the locations where users have been active. Thus, the destinations will always be used to correct for multiple counts and excess trips contained in the origins (origins from the HPP dataset) and OD pairs (from HPP dataset).

4.3 The process for creating the OD matrix

The process for creating the OD matrix from the viewDat database is the following. The following paragraphs discuss a solution to the ‘Less than 15’ problem and other aspects of the proposed procedure.

1. Total destinations per region are set equal to the sightings of the DPU dataset of 10:00 a.m.
2. Total origins per region are set equal to the origins/home locations of the HPP dataset of 10:00 a.m. – 12:00 p.m.

Note: Alternatively, the origins can be filled similarly to the method used in 4.3.2. This method should most accurately distribute the origins and counter the problem mentioned in paragraph 3 of issue (c) of subchapter 3.5. Nevertheless, if origins are calculated this way, they are no longer an observed value.

3. The total origins are equilibrated to total destinations. The procedure is similar to the method of “trip balancing” (Bovy, et al., 2006; De Dios Ortúzar & Willumsen, 2011).
4. A subscript to correct the rounding procedure is applied.
5. The OD pairs are filled from the HPP dataset of 10:00 a.m. – 12:00 p.m.
6. The OD pairs are equilibrated to the total destinations and origins using a method similar to the “balancing factors” (De Dios Ortúzar & Willumsen, 2011).
7. The ‘Less than 15’ problem can be solved given the solution in paragraph 4.3.2.

Note: Given the absence of the generalized cost values, this step is not implemented.
Algorithm: Create OD matrix from HPP and DPU datasets

Input: Raw HPP dataset, raw DPU dataset, Date
Output: OD matrix

1. DELETE entries when HPP, DPU == 9999 \| HPP, DPU == NaN \| 10000 <= HPP, DPU <= 20000 \| HPP, DPU != Date
2. DELETE entries when HPP != 10-12h and when DPU != 10h
3. SET OD(Origin, Destination) ← 0 with size equal to the unique values of HPP(BAG regions)
4. OD(Total Destinations) ← DPU /Total Destinations refer to a single region
5. OD(Total Origins) ← HPP(Home)
6. for i = 1, 2,..., size(OD)
   /loop by row to equilibrate total origins to total destinations
   7. OD ← OD(Total Origin) * sum(OD(Total Destinations)) / sum(OD(Total Origins))
8. ROUND the OD matrix
9. case D = OD(Total Origins) – OD(Total Destinations) of
22:           end for 
23:           f(iteration) = [mean(frow), mean(fcolumn)] 
24: end while 
25: RETURN OD

4.3.1 The equilibration procedure of OD pairs to total origins/destinations

After inserting the OD pairs in the OD matrix, the problem that arises is that the summation per origin (row) and destination (column), do not match to the total origins and destinations that were calculated beforehand. The procedure of equilibrating these values is incorporated in the gravity model of the 4-step model. The procedure used here is similar if not identical, to balancing factors.

The procedure starts with the rows, calculates the factor and multiplies it with the OD pairs. The same applies for columns. The whole process will have to be repeated until convergence. The factors for row (index 1) and column (index 2), are the following:

\[
(f_1)_{vi} = \left( \frac{T_i}{\sum_i T_{ij}} \right)_{vi} \\
(f_2)_{vj} = \left( \frac{T_j}{\sum_j T_{ij}} \right)_{vj}
\]

where, \( T \) are trips, \( i \) is the index of origins and \( j \) is the index of destinations. Convergence for this problem is set by a theoretical threshold (1st convergence criterion) and a practical (2nd convergence criterion). The practical threshold sets the maximum number of iterations that the algorithm can make in order to avoid an infinite loop. The theoretical limit is the average of the
difference of two consecutive calculations of $f_1$ and $f_2$. If the average is equal to 0 then the algorithm stops and the final OD matrix has been produced.

The Matlab script that was created can repeat this procedure for 5 working days starting with Monday and produce an average of the 5 OD matrices. The advantage of this procedure is that it may capture some of the regions that generated or attracted no trips due to the ‘Less than 15’ problem. However, if this averaging takes place, the solution for the ‘Less than 15’ problem must be applied before summing and averaging the matrices.

4.3.2 A solution to the ‘Less than 15’ problem

The ‘Less than 15’ problem is a major issue and can result in great loss of information. For example, in the first Monday of September 2013, there are 1,795,131 frequent visitors (0-24h period) in total of whom, 198,210 frequent visitors are registered to arrive from an ‘unknown origin’. For all months of 2013 and for all periods, there are 3,894,320 entries of ‘unknown origin’ out of 68,477,617 total entries in the dataset.

This problem can be solved using transport modelling and more specifically, a singly constrained, bound to destination, gravity model. The singly constrained is preferred over the doubly constrained because in this case the destinations are known. Singly constrained models, bound to origin, are preferred for all other purposes besides work, e.g., shopping and leisure, because for these purposes, it is hard to accurately determine the destinations (De Dios Ortúzar & Willumsen, 2011).

The formulation of the gravity model is the following:

$$T_{ij} = A_i O_i B_j D_j f(c_{ij})$$  \hspace{1cm} (6)

where, $O_i$ and $D_j$ are the origins and destinations respectively; $A_i$ and $B_j$ are balancing factors and $f(c_{ij})$ is the distribution/deterrence function which will be analysed later in this paragraph. In a singly constrained model, the factor $A_i$ is set equal to 1 and the following formula holds for $B_j$:

$$B_j = 1/ \sum_i O_i f(c_{ij})$$  \hspace{1cm} (7)

The deterrence function shows the willingness to make a trip depending on the generalized cost of the travel. It is derived from the Empirical Trip Length Distributions–ETLD of the Monitoring tool 2009. Since there is no information regarding the mode, the Trip Length Distribution–TLD for the total trips was used.

It should be noted that the ETLD provides information about trips while the GSM data provide information about users. Each trip from ‘unknown origin’ in the GSM dataset is a unique trip of the user which consequently means that a user from an ‘unknown origin’ will not make more than 1 trip. All values of the ETLD should therefore be reduced so that the total average is equal to 1. This is done proportionally by dividing each value of the ETLD per kilometre bin with 2.99. That way, the trip proportions per kilometre are preserved. These values were then multiplied with the trips of unknown origin from the GSM data. The values used in this case are for the period 0-24h for Monday 1st of July 2013. Finally, it was noticed from the distance table that there were not any BAG regions that abstain more than 0.5km and less than 1km. This would result in a very high, unrealistic factor for that kilometre bin. Therefore, the categories 0-0.5km and 0.5-1km were merged.

The estimation of the deterrence function from the empirical trip length distribution can be done using the Hyman’s or Poisson method. The former method requires that both productions and attractions are known for the trip generation step which is not the case in this problem. Poisson method relaxes this requirement by allowing for partial/observed OD matrices as input. The Poisson
method is proposed for this problem because of its versatility. The inputs of the Poisson model are the distance matrix which can be obtained with GIS and the (partial) OD matrix.

In order to estimate the deterrence function, the origins of the OD matrix were filled from the destinations using factors from the base matrix of the LMS. The restriction for the origins is that $\sum_{i=1}^{l} O_i = \sum_{j=1}^{d} D_j$. The formula for calculating the factor for each origin zone is the following:

$$f_P_i = D_j^{LMS} \cdot \frac{\text{Population}_{BAG_i}}{\sum D_j^{LMS}} \quad (8)$$

The factor in essence distributes the destinations over the origins according to the population of the zone. This was implemented because it was observed that some LMS zones were constituted by more than one BAG zone. When there were more than one LMS zones per BAG zone, the fraction of the population was forced to 1 and the destinations for the BAG region were found by summing the destinations of the LMS zones that constituted the BAG region. This is formally illustrated in (9) where: $rac{\text{Population}_{BAG_i}}{\text{Population}_{LMS_k}} = K_i$.

$$f_P_i = \begin{cases} 
T_{ij}^{LMS} \cdot \frac{K_i}{\sum T_{ij}^{LMS}}, & \text{if } BAG_i \subset LMS_k \Rightarrow i < k \\
T_{ij}^{LMS}, & \text{if } BAG_i \supset LMS_k \Rightarrow i \geq k 
\end{cases} \quad (9)$$

The following relation is applied to calculate the trips per origin:

$$P_i = \begin{cases} 
\sum_j f_P_i \cdot D_j, & \text{if } BAG_i \subset LMS_k \\
\sum_i \left( \sum_j f_P_i \cdot D_j \right), & \text{if } BAG_i \supset LMS_k 
\end{cases} \quad (10)$$

The Poisson’s method was implemented using a custom made Matlab script. The algorithm converged after 223 iterations. It was noticed that the factor for the kilometre bin 1-2.5km was too high compared to the rest resulting in a bad estimation of the deterrence function. For that purpose, the bins were aggregated to the following kilometre categories: 0-1km, 1-3.7km, 3.7-7.5km, 7.5-15km, 15-30km, 30-50km, >50km. The algorithm converged after 43 iterations (68sec on Windows 7 64bit, Intel Core i7-3610QM CPU 4 cores 2.3GHz and 5,9GB RAM).

To obtain the final deterrence function, the data were tested for fit to the following models: exponential, lognormal, power, top-exponential and top-lognormal. The best fit, based on the adjusted $R^2$ and Root Mean Square Deviation, was with the top-lognormal model. The fit was achieved using Non Linear Least Squares method. The estimated model is the following:

$$F_{ij}(c_{ij}) = 1.304 \cdot c_{ij}^{0.5271} \cdot e^{\left(-1.124 + \left(\log(c_{ij}+1)\right)\right)^2} \quad (11)$$

The problem with this model is that it overestimates trips between zones that are close to each other, e.g., neighbours, and underestimates trips that are further away (approximately more than 30km away). To solve the under/over-estimation, the power function can be used:

$$F_{ij}(c_{ij}) = 1.447 \cdot c_{ij}^{-0.226} \quad (12)$$

The power function overestimates trips that are close to the destination (first kilometre bin) but then reduces rapidly. In the case of the problem at hand, the intra-zonal trips should be forced to zero.
thus the overestimation should not damage the estimation procedure. The fit for the power function was achieved using robust least squares with Bi-Square estimators.

Since there will be no mode distinction, the cost that will be inputted into the deterrence function should be generalized cost over all modes. After computing the generalized cost per mode, the following function should be applied:

\[ c_{ij} = \frac{1}{\sum k c_{ijk}} \]  

where, \( k \) is the mode. This function represents the electric circuit analogy.

There are regions among which there already is an amount of trips. These are trips whose origin is known from the HPP dataset. Therefore, the model analysed here should not generate any trips between these regions. The only possible way to include this to the model is by forcing the values of the generalized costs of these trip combinations to be zero. All intra-zonal values should also be equal to zero.

The final issue to be examined is when the model distributes more than 15 trips to some trip combinations. This can be solved iteratively by determining the combinations that have more than 15 trips and re-assigning the excess trips using the same model while keeping the generalized costs of the regions where more than 15 trips were observed, equal to zero. The matrices are summed per iteration and the procedure is repeated.

### 4.4 Accuracy of the generated OD matrix

#### 4.4.1 Comparing the Trip Length Distributions

The observed trip length distribution (ETLD) is taken from Monitoringstool 2009. The data show how many trips a person made in a day for both hours of morning peak, 8:00-9:00 a.m. and 9:00-10:00 a.m.

![Figure 26: Comparison of the trip length distributions of the Monitoringstool 2009 and the OD from the viewDat database](image)

**Figure 26: Comparison of the trip length distributions of the Monitoringstool 2009 and the OD from the viewDat database**
It seems that the trip length distribution after 7.5 km is similar for both datasets. However, the problem of the spatial aggregation, whose effects can be seen at kilometre bins 0.1-3.7, clearly damages the quality of the data. The BAG regions of the 10 biggest cities in the Netherlands have an average size of 91.21 km² with a minimum value of 31.01 km² for Haarlem and a maximum value of 196.62 km² for Amsterdam. For Monday 3rd of March 2014 the amount of intra-zonal trips was calculated to be 778,850 trips or 61% out of the 62% of the total trips that appear in the first kilometre bin of Figure 26. This seems to be the case for all other days as well. Figure 27 shows the unaltered TLDs of OD matrices that are calculated for the other working days of the first week of March.

![Figure 27: The Trip Length Distribution for OD matrices of all the working days of the first week of March](image)

To verify that this phenomenon is due to spatial aggregation, the trips of the first kilometre bin are redistributed to other kilometre bins. If the problem is spatial aggregation, then the distribution should reach the same result as the ETLD. It can be that the absolute values of each kilometre bin end up higher or lower than ETLD. Then the problem should lie in the method of deriving the OD or the original estimation of sightings of the HPP and DPU datasets.

The zones of the largest cities of The Netherlands should affect the kilometre bins up to 7.5km. Trips until 3.7km are not corresponding to the observed values but only values that are above 7.5 km show similarities. In between these two limits it is uncertain whether the comparison shows similarities or not (kilometre bins 3.7-5 km and 5-7.5 km). Thus, two scenarios are created; the trips of the first kilometre bin will be distributed in the kilometre bins until 5 km and until 7.5 km.

The amount of trips in the first bin is: 0.62 * 1,268,542 = 786,496 ‘undistributed trips’. Distribution of trips was performed using the proportions of the ETLD. The results of the two scenarios are shown in Figure 28. Additionally in Figure 27, the unaltered TLDs of OD matrices calculated for the other working days of the first week of March are shown.
The action of distributing the "undistributed trips" seems to improve the overall picture of the trip length distribution using either scenario. The fact that the second scenario results in the overestimation of trips in the 5-7.5 kilometre bin is of note. In general, the process above shows that the derived OD miscalculates the trips until 5km but certainly below 7.5km. The initial hypothesis, especially for the first scenario appearing in grey in Figure 28, is correct. The "undistributed trips" of the first kilometre bin seen in Figure 26 are actually trips of longer distance that were regarded as intra-zonal. Another argument that supports this idea is that most of these trips have taken place with slow modes being moped, bicycle and walking, modes that are used inside the cities and towns and for short distances.

4.4.2 Sparseness of the matrix

Another way to determine matrix accuracy is to determine the number of 0 values. When there are 0 trips between regions then this indicates that these regions have no relationship or that they are not generating or attracting trips between each other. The amount of zeroes is called sparseness of the OD matrix (Ton & Hensher, 2002; De Dios Ortúzar & Willumsen, 2011). Sparseness can be attributed to sample size, geographical specification of the traffic zones and disaggregation with respect to trip purposes (Ton & Hensher, 2002). Since there is no division to trip purposes and the sample size should be large enough to affect sparseness only positively, the negative effects should be (mostly but not only) attributed to geographical specification of BAG regions. An additional problem that can affect the sparseness of the OD is the positioning of mobile phones to regions. It should also be noted that Ton & Hensher (2002) encountered the exact same problem with the ‘Less than 15’ problem. In their case, measurements below 200 trips were set to 0 due to privacy issues. They also reported that the matrix had 67% of its cells empty referring to a matrix for all modes, trip purposes and times of day.

For the generated matrix from the viewDat database for morning peak of Monday 3rd of March 2014, 99% of the cells are empty which means that only 12,214 cells out of 1,315,609 cells have a value different than 0. The process was repeated for every day of the first week of March 2014 and the percentage of cells with 0 values over the total amount of cells was always equal to 99%. This is a high value but is not necessarily incorrect. For comparison, the sparseness of the LMS matrices for car was calculated and is shown in Table 9. The sparseness when removing values

![Figure 28: Distribution of the "undistributed trips" over the kilometre bins](image-url)
below 15 which is equivalent to the ‘Less than 15’ problem is also shown below. The total cells of the LMS matrix are 2,365,444 (1538 zones).

Table 9: Sparseness of the LMS matrices

<table>
<thead>
<tr>
<th>Calculated for</th>
<th>Time of day</th>
<th>Level of sparseness</th>
<th>Amount of non-empty cells</th>
<th>Total Trips</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater than 0 (&gt;0)</td>
<td>Morning peak</td>
<td>92.75%</td>
<td>171,439</td>
<td>1,301,008</td>
</tr>
<tr>
<td></td>
<td>Evening peak</td>
<td>92.58%</td>
<td>175,550</td>
<td>1,718,107</td>
</tr>
<tr>
<td></td>
<td>Rest of day</td>
<td>93.93%</td>
<td>143,527</td>
<td>888,906</td>
</tr>
<tr>
<td>Greater than 15 (&gt;15)</td>
<td>Morning peak</td>
<td>99.39%</td>
<td>14,501</td>
<td>857,245</td>
</tr>
<tr>
<td></td>
<td>Evening peak</td>
<td>99.26%</td>
<td>17,545</td>
<td>1,259,290</td>
</tr>
<tr>
<td></td>
<td>Rest of day</td>
<td>99.62%</td>
<td>9,100</td>
<td>533,575</td>
</tr>
</tbody>
</table>

Table 9 shows that the sparseness of the matrices for the derived OD and the LMS (for the cells that are greater than 15) are similar. However, this comparison refers to the matrices as a whole. It could be that different cells are not zero at each matrix. Direct comparison of the matrices is impossible because the spatial specification of the two matrices do not match. Nevertheless, there can be a correspondence if both matrices are transferred to a different, yet similar, specification. The specification that can be used as a reference for both matrices is a simple square grid which can be seen in Figure 29.

The grid is created from performing (centroidal) Voronoi tessellation (Du, et al., 1999) on a set of regular points. The reference system used is Amersfoort / RD New / EPSG: 28992, which explains the bent figure of the grid of Figure 29. Using the intersection tool of QGIS, it is possible to determine which centroids of BAG and LMS zones are included in each rectangle of the grid. Then, all trips of the zones are transferred to a rectangle of the grid.

All rectangles are equal and constrained by the size of the underlying layer. The size of the rectangles was chosen so that either no centroid of LMS or BAG regions is included in the rectangle or at least one centroid of a BAG region and one centroid of an LMS zone are included. This rule results in non-unique solutions, thus, the algorithm was set to look for the solution that provided the closest result (the less amount of regions with only one centroid, either from LMS or BAG). The result is approximately a 10x10 square grid.
All zones, for both LMS and BAG, are transformed to the new zoning system. The comparison of the two matrices for sparseness should incorporate the spatial characteristics of the matrices. The spatial information or spatial correlation in the matrices is the structural information of the matrices since this information gives the matrix its form. A new method to determine the differences between two matrices while accounting for the spatial correlation enclosed in their structure is the Structural SIMilarity index or SSIM.

The SSIM was introduced by Wang et al. (2004) as a tool for image comparison that accounts for the structure of images. The structure of images is important to be compared, in order to determine image similarity, because the pixels of images that are close to each other appear to have interdependencies. These interdependencies are much similar to the spatial correlation of regions in an OD matrix. Djukic et al. (2013) showed that contextually and mathematically, the SSIM is applicable to OD matrices and is conceptually more effective than other measures that do not account for spatial correlation such as the Mean Squared Error.

The formula used for applying the SSIM is:

$$SSIM(\text{LMS, GSM}) = \frac{2 \cdot \mu_{LMS} \cdot \mu_{GSM} + C_1}{\mu_{LMS}^2 + \mu_{GSM}^2 + C_1} + \frac{2 \cdot \sigma_{LMS-GSM} + C_2}{\sigma_{LMS}^2 + \sigma_{GSM}^2 + C_2}$$

(14)

where, $\mu$ is the mean of each dataset; $\sigma$ is the variance of each dataset; $\sigma_{LMS-GSM}$ is the covariance of the two matrices; $C_1$ and $C_2$ are stabilization coefficients.

The SSIM is calculated according to the methodology described by Wang et al. (2004) and Pollard et al. (2013) which is on a frame of the matrix. There is no indication of the appropriate size of frame in current literature. Therefore, a 10x10 frame will be used. Using an iterative procedure, the SSIM is calculated per 10x10 overlapping squares moving one column to the right per iteration. When the end column of the matrix is reached, the SSIM is calculated for a 10x10 matrix whose first column is the first column of the OD matrix but whose first row is the next row than previously. The procedure recalculates the SSIM per 10x10 matrices moving one column per iteration and repeats the procedure until the end column and end row are checked. The final index is the average of the calculated SSIMs and is defined as follows (Pollard, et al., 2013):

$$MSSIM = \frac{1}{N} \sum_{i=1}^{N} SSIM_N$$

(15)

where, $N$ is the number of 10x10 matrices that were used in total to calculate the SSIM. MSSIM stands for Mean Structural SIMilarity index. The index takes values between -1 and 1. A value of 0 would mean that the matrices are completely different while a value of 1 would mean that the matrices are identical.

Using SSIM for comparison of the sparseness of two matrices require that the absolute values of trips do not damage the estimation procedure. Therefore all non-zero cells of both matrices are set equal to 1. The calculation of the MSSIM gave a result of 0.5854. This result shows that there are significant spatial differences between the way that sparseness affects each matrix and also, that there are more significant similarities. It is however certain that the sparseness is not the same for both matrices.

### 4.4.3 Examining specific regions

Analysing specific regions may give insight into the accuracy and characteristics of the OD matrix. The 5 regions that are analysed in this paragraph are Utrecht, Groningen, Eindhoven, Almere and Deventer. The following table includes the origins and destinations from the derived OD.
Chapter 4 - Acquiring Origin-Destination information from the viewDat database

Table 10: Calculated origins and destinations for selected regions

<table>
<thead>
<tr>
<th>City</th>
<th>Origins</th>
<th>Destinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utrecht</td>
<td>29,771</td>
<td>38,882</td>
</tr>
<tr>
<td>Groningen</td>
<td>19,700</td>
<td>22,652</td>
</tr>
<tr>
<td>Eindhoven</td>
<td>18,640</td>
<td>18,369</td>
</tr>
<tr>
<td>Almere</td>
<td>18,192</td>
<td>14,327</td>
</tr>
<tr>
<td>Deventer</td>
<td>5,473</td>
<td>5,434</td>
</tr>
</tbody>
</table>

From the first look on the data, the results seem reasonable. Utrecht has the higher origins/destinations values since it is the largest city of the 5. It also attracts more trips than it produces which also reasonable since it has a lot of work places and is the largest shopping centre of the Randstad. Almere on the other hand produces more trips than it attracts which can be justified because it is a satellite city of Amsterdam for work and other purposes as well.

The ratio of origins and destinations is not directly comparable to a known statistical value. It could be that it is comparable to the ratio of origins and destinations from the LMS. But, the LMS concerns only car mode. Even so, the comparison gives an indication of soundness of the calculated values. These are tabulated in Table 11.

Table 11: Ratio of origins over destinations for viewDat database and LMS

<table>
<thead>
<tr>
<th>City</th>
<th>Ratio from viewDat dat.</th>
<th>Ratio from LMS (morning peak)</th>
<th>Ratio from LMS (rest of day)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utrecht</td>
<td>77%</td>
<td>75%</td>
<td>103%</td>
</tr>
<tr>
<td>Groningen</td>
<td>87%</td>
<td>73%</td>
<td>100%</td>
</tr>
<tr>
<td>Eindhoven</td>
<td>101%</td>
<td>78%</td>
<td>102%</td>
</tr>
<tr>
<td>Almere</td>
<td>127%</td>
<td>113%</td>
<td>96%</td>
</tr>
<tr>
<td>Deventer</td>
<td>101%</td>
<td>88%</td>
<td>99%</td>
</tr>
</tbody>
</table>

Most of the multiple counts’ and excess trips’ effects should have been eradicated. Nevertheless, the fact that the ratio for the viewDat database is always higher than the ratio from the LMS indicates that this is not the case. The ratios from LMS for the ‘rest of day’ period are more balanced; origins and destinations are almost equal. Besides the multiple counts, the time interval that was used to create the OD matrix is also an issue. It is likely that because the 10-12h period was used, trips that do not belong in the peak hour are also included. This would explain why the percentages for Eindhoven and Deventer are elevated and closer to 1.

The ratio of origins and destinations over the total trips for the Netherlands from the derived OD show the share of trips for the specific region. More specifically, it shows how each region affects the total mobility of the Netherlands. The ratios are compared to their counterpart from the LMS.

Table 12: Origins and destinations per the selected regions over total trips for the viewDat database and LMS

<table>
<thead>
<tr>
<th>City</th>
<th>Ratio of origins (viewDat database)</th>
<th>Ratio of destinations (LMS)</th>
<th>Ratio of origins (viewDat database)</th>
<th>Ratio of destinations (LMS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utrecht</td>
<td>2.34%</td>
<td>1.43%</td>
<td>3.06%</td>
<td>1.91%</td>
</tr>
<tr>
<td>Groningen</td>
<td>1.55%</td>
<td>1.10%</td>
<td>1.78%</td>
<td>1.51%</td>
</tr>
<tr>
<td>Eindhoven</td>
<td>1.46%</td>
<td>1.60%</td>
<td>1.44%</td>
<td>2.04%</td>
</tr>
<tr>
<td>Almere</td>
<td>1.43%</td>
<td>1.07%</td>
<td>1.13%</td>
<td>0.95%</td>
</tr>
<tr>
<td>Deventer</td>
<td>0.43%</td>
<td>0.54%</td>
<td>0.43%</td>
<td>0.61%</td>
</tr>
</tbody>
</table>
The percentages are similar. The difference in Utrecht should be attributed to the fact that the correspondence of regions is not exact (aggregation of LMS zones is smaller than the BAG region) and to the increased use of train. Utrecht has the largest train station in the Netherlands. Thus, a number of people will use the train to reach Utrecht while train is not included in the LMS matrix.

The trip proportions between Almere and Amsterdam are also of interest. This trip end combination is special because of the relationship of the two cities. There are a number of workers of Amsterdam that live in Almere making the trip combination from Almere to Amsterdam show an increased number of commuters (Van der Worp & Beeckman, 2013). For the viewDat database, the ratio of trips from Almere to Amsterdam over the origins of Almere is equal to 10%. For LMS, the same ratio is equal to 14%. Although the percentages are similar, it was expected that the former would be higher than the latter since Public Transport – PT modes are also included. It may be that PT increases the total origins from Almere more than it increases the trips between Almere and Amsterdam. Therefore, people will use PT to travel to other destinations than Amsterdam which effectively reduces the total trips to Amsterdam in comparison to trips only with car mode. The small distance of Almere and Amsterdam (22km) also justifies higher car usage. The train is more attractive for longer distances.

The trips ratios (trips of a specific OD pair over the total productions of the origin) where examined for the other cities as well. More specifically, the two destinations that were examined are Amsterdam and The Hague. The origins are always the selected regions.

<table>
<thead>
<tr>
<th>City</th>
<th>Trips to Amsterdam (viewDat database)</th>
<th>Trips to Amsterdam (LMS)</th>
<th>Trips to Den Haag (viewDat database)</th>
<th>Trips to Den Haag (LMS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utrecht</td>
<td>5%</td>
<td>2%</td>
<td>2%</td>
<td>0.18%</td>
</tr>
<tr>
<td>Groningen</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.02%</td>
<td>0.01%</td>
</tr>
<tr>
<td>Eindhoven</td>
<td>2%</td>
<td>0.1%</td>
<td>1%</td>
<td>0.04%</td>
</tr>
<tr>
<td>Almere</td>
<td>10%</td>
<td>14%</td>
<td>0.4%</td>
<td>0.11%</td>
</tr>
<tr>
<td>Deventer</td>
<td>0.4%</td>
<td>1%</td>
<td>0.1%</td>
<td>0.11%</td>
</tr>
</tbody>
</table>

Finally, the trips from the derived OD matrix were compared to the commuters’ (forenzen) origin-destination data (Van der Worp & Beeckman, 2013). The comparison was done for the 5 cities used above and Amsterdam, Rotterdam, Den Haag, Tilburg, Breda and Nijmegen. The BAG regions were aggregated appropriately to match the specification used by CBS. Scaling to the population was done by multiplying the derived OD matrix with the fraction of the maximum values of the two matrices, which are the intra-zonal trips of Amsterdam for both matrices. The following observations have been made:

- Mean value of the commuter matrix is 9878 while the mean value of the derived OD is 9564. This shows that the total trips in each matrix are comparable.
- The matrices were compared using a heat map. The heat map shows the results of the subtraction of the two matrices in percentage values (division by the summation of the values of the cells). Figure 30 showed that there are not significant differences between the two matrices besides in 4 out of 121 specific occasions/OD pair values.

The tests showed with good confidence that the derived OD matrix contains trips that are in agreement with the observations made by multiple different sources.
4.5 Conclusions about the accuracy of the assumptions

The accuracy of the derived OD matrix resides on the truthfulness and power of the assumptions that were made in subchapter 4.2. Although, the conclusions were embedded in the text of the previous paragraphs, this subchapter discusses the results of the tests for the accuracy of the derived OD matrix with respect to each of the assumption made.

■ Assumption 1: The DPU sightings per region are the total destinations per region of the OD matrix
Verifying this assumption requires the examination of the destinations of the derived OD matrix. First of all, the matrix that is made is a morning peak matrix. This was shown in Table 11 where the ratio of origins and destinations is not balanced. Some regions act as attractors and others as generators of trips. This is congruent with the corresponding ratio from the LMS matrix. The destinations also seemed to be comparable to LMS per region as Table 12 shows. Ordering the values in descending order from LMS shows that Eindhoven is the greatest attractor in comparison to the derived OD that shows Utrecht as the greatest attractor. Nevertheless, the LMS only contains trips by car while Utrecht has the largest train station in the Netherlands. Thus, it is more accessible by train than Eindhoven, something that is not obvious in the LMS matrix but can be hypothesized from the derived OD matrix. Furthermore, the TLD showed that the length of trips is comparable to the observed values of the Monitoringstool which enhances the argument that the distance between the origins and destinations in the derived OD are tangent to reality. Finally, the fact that the trips in the first kilometre bin of the TLD could be distributed to the kilometre bins that may have fallen within the borders of some major BAG regions, shows that the intra-zonal trips of the derived OD are also consistent to observations had it not been for the spatial aggregation.

■ Assumption 2: The total origins per region are drawn from the HPP dataset for the 10:00a.m.-12:00p.m. period
The same reasoning as above applies for this assumption. However, the initial values of the matrix prior to equilibrating total origins and destinations, shows that the amount of origins is much higher than the destinations. Therefore, problem 3 of issue with index (c) of subchapter 3.5 may result in increased number of origins that damages the ratio of origins. Table 11 also
showed an increased number of origins since the ratio of origins over destinations is consistently higher than the equivalent in LMS. Solving this issue seems impossible because it would require assumptions on the way that origins are generating trips. A proposal is done in paragraph 4.6 that suggests an alteration in the specification of the HPP dataset in order to account for the increased number of origins that was encountered.

**Assumption 3: The origin-destination combinations of the HPP dataset contain accurate enough spatial information to be used as the OD pairs’ values**

Examining the sparseness of the matrix showed that both matrices are equally sparse. When examining the amount of cells that contain below 15 trips in the LMS matrix, it was observed that both the LMS and the derived OD matrices contained empty cells in almost equal amount. However, this did not prove whether the cells that are 0 in the derived OD matrix are also 0 in the LMS matrix (or below 15 to be exact). After careful examination it was shown that the sparseness is similarly distributed for both matrices at a great extent. However, the comparison also showed that there are numerous 0 cells in one matrix that do not correspond to the other. Considering the differences in the specification as far as mode is concerned, the results is not inconclusive, the matrices showed good correspondence in the degree and distribution of their sparseness. Nevertheless, further and thorough examination is required preferably by comparing to other sources such as OV-chipkaart data (in order to lessen the extent of the difference in mode).

A final remark on the estimation procedure concerns the assumption that destinations are more accurate than the origins. The inconsistency found in the origins shows that this assumption was well-founded and provided that the following proposal is implemented, this assumption will be rendered irrelevant.

### 4.6 Recommendations

**4.6.1  An improved specification of the HPP dataset**

The specification of the HPP dataset is problematic. The two biggest issues with its specification are first, the multiple counts/excess trips of the same device/individual and second, the temporal aggregation of the HPP dataset.

Improving the specification means that the first two problems will be abolished or their effects will be minimized. The solution should take into consideration the final result which is to derive trip information. Therefore there should be a clearly defined origin and a clearly defined destination. Because Mezuro has developed an algorithm for determining home locations, it would be unwise to request to change it instead of improving it.

In transportation modelling and planning studies, the home is the predominant origin in the early hours of the day. Any trip with the home as origin should begin in the early hours of the day which is until 10:00 a.m. This does not exclude the home being an origin at other hours of the day or does not insinuate that home is the only origin in the morning peak hours. But, it was mentioned in 4.3.2 that trips with home as origin, dominate the morning peak hour travel patterns.

The concept above justifies the use of home as an origin. Solving the multiple counts and improving the general quality of the HPP dataset could be achieved by implementing one of the following recommendations:

1. Only the last location of the user should be logged. There does not seem to be a valid reason why all events will be logged into the HPP dataset besides that it is computationally easier. Mezuro has access to the unique ID’s of each mobile phone which means that they can dissociate each user from other users. Consequently, it would be more useful to show the last known location of the user instead of all his/her locations.
2. The most active location should be shown. The DPU dataset uses the exact same concept: the user is logged at the location where he/she made most events in the specified time interval. Since the algorithm for determining such locations is already developed by Mezuro, it could be easily transferred and applied on the HPP dataset.

From the two aforementioned recommendations, the second should be preferred because it includes an indication for identifying activities which is a user remaining in a location for some time. It is important to note that ‘excess trips’ will also be excluded using any of the aforementioned recommendations. But, using the 10-12h period as the pool where draw the origins from, is still problematic. The reason is that, in case of the first recommendation, the user might be located in the destination of a trip with purpose other than work while in the case of the second recommendation it is uncertain where the user will be logged.

This can be countered by using a time period that is closer to the morning peak such as 8-10h, 8:30-10:30h or 9-11h. In the case that the user is still travelling in the morning peak, then he/she will not be logged in the place where he/she was seen the most because there will be no such place. Logging the last location would be more efficient in earlier time periods being 8-10h. Summarizing, in case that the periods are not changed (or the 9-11h period is used), the second recommendation (most active location) should be preferred. If the periods can be altered then the first recommendation (last location) should be implemented for the 8-10h period.

The recommendations above will improve the quality of the HPP dataset without altering the current uses of the dataset. This is argued because currently the HPP dataset shows the traces of users. Thus, it doesn’t include all the regions the user went through, because he/she doesn’t necessary make an event in every region, and there is no way to determine the crossed regions (without stop) versus the regions where the user stopped. In other words, the traces are fragmented dependent on phone usage.

Applying these improvements does not require the methodology or algorithm described in chapter 4 to be altered. The change will improve the estimation procedure immensely making the origins of the matrix more trustworthy.

4.6.2 The SSIM with low variance and/or covariance

SSIM is not accurate if the variance of the matrices is low or more specifically, below 1. This can be improved by replacing factor \(C_2\) with 1. This formulation allows comparison of matrices that are, e.g., in percentiles. This change is not important if both the nominator and denominator of the second fraction of formula (16) are greater than 1 but the formula should collapse to the original formulation if the nominator and denominator are close but greater than 1. Any mixed case of the nominator and the denominator should follow the formulation above. The proposed formulation for the SSIM for comparing matrices would be:

\[
SSIM(M_1, M_2) = \begin{cases} 
S_1 \frac{2 \cdot \sigma_{\text{M1,M2}} + 1}{\sigma_{\text{M1}}^2 + \sigma_{\text{M2}}^2 + 1}, & \text{if } \sigma_{\text{M1,M2}} < 1 \text{ or } \sigma_{\text{M1}}^2 < 1 \text{ or } \sigma_{\text{M2}}^2 < 1 \\
S_1 \frac{2 \cdot \sigma_{\text{M1,M2}} + 1 + C_2}{\sigma_{\text{M1}}^2 + \sigma_{\text{M2}}^2 + C_2}, & \text{otherwise}
\end{cases}
\]

\[S_1 = \frac{2 \cdot \mu_{\text{LMS}} \cdot \mu_{\text{GSM}} + C_1}{\sigma_{\text{LMS}}^2 + \sigma_{\text{GSM}}^2 + C_1}\]
5. Additional methods for extracting OD matrices from aggregated data

5.1 Initial remarks on aggregated GSM datasets

There are no papers to this date describing methods to extract origin-destination data from aggregated GSM data. Researchers claim that there cannot be adequate disaggregation to trip purposes or modes and that in general, the aggregated data do not allow the study of individuals (Calabrese, 2011; Wolf, et al., 2014b). Studying individuals allows for employing stop and go algorithms that can determine if the user has stopped to participate in an activity or not. Without this vital information, determining trips is next to impossible. However, due to strict privacy concerns, obtaining databases with individual information might also be hard. The privacy concerns do not only extend to the information that are commercially available, for the Netherlands being aggregated data instead of individual, but also to the way that the data are mined.

The mobile network operator does not allow the raw data to be stored locally but that all the necessary operations on the data take place on the mobile network operator’s servers and performed remotely. This creates problems with the communication because the speed of processing and of data exchange is bound to the speed that the mobile network operator’s servers allow. Mezuro can perform queries on the data and only extract the results, not the data themselves. An additional problem occurs because the mobile network operator allows the queries to be run for a month and an additional 5 days. After that timeframe, the data are purged from the database.

Various commercially available GSM and other databases with location information for transport use may be derived from similar processes. A study by Hoh et al. (2006) showed that a good practice for sellers of transport data is to split the data server from the transport server and ensure that the information coming out of the data server are anonymized and cannot be used to trace individuals. Thus, to ensure privacy concerns, it is easier that databases become aggregated, similarly to what was already seen with viewDat database (Wolf, et al., 2014a). This thesis up to now engaged the issue of deriving Origin-Destination matrices from a very restricted environment as far as available information goes. Determining Origin-Destination matrices is a process governed by uncertainty because conclusions need to be derived about multiple choices of travellers, i.e., destination, purpose, time of day, mode and route. Decreasing the available information seems
counter-intuitive but extremely pragmatical given the circumstances and the general trend of the market which is to ensure the privacy of users (Wolf, et al., 2014a; Wolf, et al., 2014b).

Creating databases that are more aggregated or are less related to users might solve the major privacy concerns which are positioning of individual users and their identification. This chapter contains two methodologies that could work with databases that do not contain direct information on movements of individuals or groups of individuals.

5.2 Possibilities for deriving OD information from aggregated GSM data

The sightings in any dataset of aggregated GSM data including the viewDat database can be thought of as created from two different categories of users: first, the ‘passer-by’ who is the user that appears only once in a region because he did not make a stop in that region and second, the ‘arriving’ user who makes multiple events in a region because he stopped at that region to participate in an activity. Given these definitions, the problem can be more accurately articulated by saying that the passers-by need to be removed from the total sightings in order for only ‘arriving’ users to be determined.

An additional problem can be created due to users not making events or making events in such a way that can only be categorized as ‘passers-by’. This can happen because the difference in the two categories is essentially the amount of events created per region. This could be solved using a statistical method that takes into consideration the amount of events per user (see Appendix B). It would be even better to exclude users from the sample that do not meet some criteria. Thus, create a smaller sample of users that are active enough to be categorized as either ‘passers-by’ or ‘arriving’. This would annul a portion of the effects mentioned by Wang et al. (2014) that active users appear as more mobile (see subchapter 2.6). This technique has been already used in GSM transportation studies (Frias-Martinez, et al., 2010; Calabrese, et al., 2011; Smoreda, et al., 2013).

5.2.1 A data-focused method for excluding ‘passers-by’

This option for creating OD matrices from aggregated data uses the same datasets as in the viewDat, meaning the HPP and DPU datasets, with the difference that the DPU dataset is a dataset of events. Essentially, the required event dataset is an even simpler dataset than the one in viewDat database. The required inputs are the following:

1. An empirical trip length distribution.
2. The spatial resolution (GIS maps) of any TAZ (Traffic Analysis Zone) such as LMS and NVM regions or even the BAG regions.
3. An event database (name in this paragraph DPUE) and a database identical in specification to the HPP (named in this paragraph HPP).
4. Data for the amount of active users. Active users are the amount of phones that created an event.

First point of note is the difference of the HPP and the DPUE databases. In the DPUE database, if a user makes more than one event in a day, he appears as many times as the events he/she did. The HPP database on the other hand, shows users: whenever a user crosses a region, he increases the number of total user sightings by 1. The user might change regions in the same period and make an event at both regions. In this case he will increase the number of sightings by 1 in both regions. As mentioned before, the sightings of the HPP database per region are the traces of the users’ movement in the spatial plane. Therefore, the HPP database, does show users; it is
definite that one sighting is one user and that the number of sightings is the exact number of users’ sightings per region they crossed.

Consider 100 travellers crossing a region without making a stop to participate in an activity while making a single event in that region. Then the number of sightings in the DPUE and the HPP database would be exactly the same and equal to 100. In this case, the number of sightings of the DPUE database is summed over a whole day. The amount of users of the HPP database is summed for all origins. In the common case where some users remain in the area and others only pass by, the numbers of sightings for the two databases will not be the same. The number of sightings in the HPP database will be 100 because 100 users created an event in the zone. The number of sightings in the DPUE database on the other hand will be higher than 100 because all users crossed the region while some of them remained, creating more events in time. The difference between the two amounts of HPP and DPUE will then be equal to the events that were created by the ‘arriving’ users, excluding the users that were only passing-by.

Figure 31: How passers-by and arriving users appear in the in the HPP and DPUE datasets

The events per user per day must also be calculated. This can be done by dividing the total sightings in the DPUE database, which are the events, to the active users per day. In the specific case of the viewDat database, the data for active users are given for the complete dataset -there is no spatial resolution- therefore, one cannot make a spatial distinction of the events per user per day using the active users’ data. Changes in the phone usage per region should be expected because not all regions are the same. Some regions are mostly residential others are mostly commercial while most are mixed. A person that is in a strictly residential region would create fewer events in a day because there is a reduced activity of phones when somebody is at home than when he is not (Patel, et al., 2006). Additionally, studies with GSM data usually have datasets from a specific provider whose share over the market might change. This was noticed for the viewDat database and discussed in chapter 3.

However, it is possible to include the spatial effects on mobile phone usage using the following formula. For a more complete explanation of the formula see subchapter 4.3.2.

\[
\rho_{region} = \frac{\text{Events}_{region}}{\text{Population}_{region}} = \frac{\text{Events}_{total}}{\text{Population}_{total}} = \frac{\text{DPUE}_{region}}{\text{DPUE}_{total}}
\]

Up to now, the two amounts that are calculated are the following:

\[
\frac{\text{DPUE}_{total}}{\text{Active users}} \cdot \rho_{region} = \frac{\text{events}}{\text{user}} \cdot \text{per day and per region}
\]

\[
\text{DPUE}_{region} - \text{HPP}_{region} = \text{events of 'arriving' users per day and per region}
\]
Dividing the aforementioned values will result in determining the ‘arriving’ users which are the users that remained in an area and were not only ‘passers-by’. Therefore the result is the number of users whose destination is the examined region. The number of origins is known from the HPP database per region (given the constraints discussed in chapter 4). At this point, the total destinations per region are known but not the trips per OD pair. Additionally, there is no guarantee that origins are equal to destinations. If this is the case, the matrix can be adjusted using the method discussed in paragraph 4.3.1 and explained by Ortúzar & Willumsen (2011). The origins and destinations after trip balancing are called ‘calculated’.

The trips can be distributed per OD pair using different methods. Three possible solutions are formulated below:

a) Simple factor of “trips” to a single destination over the total incoming trips of that destination:

\[
T_{ij} = \frac{HPP_{ij}}{\sum \limits_i HPP_{ij}} \cdot D_j
\]

where, \(i\) and \(j\) is the index for origins and destinations respectively and \(D\) is the amount of destinations which is the amount of ‘arriving’ users.

The problem with this factor is that it eventually distributes trips based on a “sighting” value and not on an actual trip value. Regions with a lot of passers-by will be grossly overestimated. Especially trips between regions that are close to each other will be increased because most travellers will cross these regions to move to their final destination.

Correcting for the aforementioned problem can be achieved by understanding that the major issue in this case is the distance between regions. There is no other effect, besides phone usage, that can result in overestimation (or underestimation) of trips between zones. Additionally, in this problem, the phone usage was inserted into the calculation procedure previously. Therefore, most of its effects should have already been alleviated.

Solving for the overestimation of trips can be done by correcting the trip length distribution function. The procedure is bound to be iterative:

i) Calculate the trips per OD pair using formula (18).

ii) Calculate the trip length distribution of the OD matrix. The difference of the trip length distribution of the OD matrix to the observed trip length distribution from MON 2009 (or any other source) can be used as a factor to reduce or increase the trips per OD pair and per kilometre bin. The factor is an index relative to 1 where 1 means that the trip length distributions, for the specific kilometre bin, match.

iii) The summation of the trips per destination and per origin will not necessarily match with the previously found values (the origin and destination values from this summation are called ‘new’). The matrix has to be corrected so that the total origins and destinations are equal to those found before (calculated values). Therefore, the matrix should be balanced (iteratively) using the factors:

\[
r_1 = \frac{\text{Origins}_{\text{Calculated}}}{\text{Origins}_{\text{New}}} \quad (19)
\]

\[
r_2 = \frac{\text{Destinations}_{\text{Calculated}}}{\text{Destinations}_{\text{New}}} \quad (20)
\]

iv) Repeat steps (ii) and (iii) until convergence. Convergence is defined by: \(r_1 \approx r_2 \approx 1\)

The above is a simplified methodology of the trip distribution step of the 4-step model which uses different inputs. The advantage of this method over using a gravity model or a discrete choice model is that it requires no external input besides the trip length distribution. The
gravity model requires (generalized) cost and the discrete choice model requires observed (revealed and/or stated) preference.

There is one assumption that needs to hold for the aforementioned procedure to be valid: the sample is big enough to accurately describe the mobility patterns of individuals in an aggregate level. In general, the trip length distribution has been used before to determine if the mobility patterns arising from GSM data are the same as the observed patterns (González, et al., 2008; Calabrese, et al., 2013).

An important note should be made about the trip length distribution. The empirical trip length distribution from the Monitoring tool 2009 concerns all trips made by travellers. On the other hand, the HPP dataset contains information about trips made from home alone. Therefore the trip length distributions of MON 2009 and GSM data (viewDat database) should not match by definition. There is not enough information to correct this inconsistency using empirical trip length distributions. If the trip length distributions from LMS are used, then only the mode of car will be used to form the distributions, which is also incorrect. The problem could be solved by using the trip length distribution from a model that studies all modes such as the NVM and by making the assumption that users make one trip from home (thus calculate only morning peak matrix as in chapter 4).

b) Trip proportions from the LMS or another transport planning model. The advantage is that this factor captures the spatial correlation of OD pairs and is in the unit of trips. On the other hand, the spatial resolution of LMS zones and BAG zones requires an adjustment that introduces uncertainty into the calculation and complicates it. A possible way to determine the trips using the LMS trip proportions is by using the following formula:

\[
 f_{Pi} = \frac{\left[ O_{i}^{LMS} \cdot \frac{\text{Population}_{BAG}}{\sum_{j} T_{ij}^{LMS}} \right] + \left[ D_{j}^{LMS} \cdot \frac{\text{Population}_{BAG}}{\sum_{i} T_{ij}^{LMS}} \right]}{2}
\]

where, \( O \) is the amount of origins.

This formula is a coarse approximation. It considers that both origin and destination contribute equally and in the same way to trip distribution. Additionally, according to the formula, the correspondence of trips among LMS zones against BAG zones is only based on population which produces a correct estimator probably only for origins. Not knowing the purpose of travel, it is impossible to improve the destinations’ part of the factor without making an assumption. Furthermore, the LMS concerns only car and freight on the Base Matrix level while the GSM data concern all modes and there can be no distinction among them. The NVM for all modes should provide a better result.

c) Usage of a trip distribution model. This will definitely be the most accurate solution since it introduces modelling and transport theory into the procedure. However, it is a data demanding method while the same adjustment, for trips originating only from home, should be made at the trip length distribution.

The OD that is created concerns the sample of the subscribers of the mobile network operator and is not for the whole population of the Netherlands. Either the sample will have to first be raised to the population and apply the aforementioned methodology or scaling-up to population will be done afterwards but using values of trips instead of persons. The current methodology could use values of persons which is already available in the viewDat database.
Scaling-up the sample at the trip level could be done using the proportions of the trips per region of the NVM over the total trips of the NVM after summing all modes. The proposed factor for scaling up is the following:

\[ f_3 = p_i^{NVM} \times \left( \frac{\text{Population}_{BAG,n}}{\text{Population}_{NVM,n}} \right) \sum_{ij} t_{ij}^{NVM} \]  

(22)

The trips are scaled up according to the origins because they are considered to be more accurately calculated versus the destinations. The final destinations are simply a summation of the corresponding trip values of OD pairs (columns) of the OD matrix.

The methodology is illustrated in Figure 32. The potential issues with the proposed methodology are the following:

a) Applying the methodology for determining trips of the ‘Less than 15’ problem, should always take place before scaling-up the sample because the results from the methodology are corresponding to the specification of the dataset. Scaling-up takes into consideration the existing sample. If the trips of ‘unknown origin’ are added later, then scaling-up will result in overestimating the persons.

b) Consider the same example of a single region that has an influx of 100 mobile phone users and that all make 1 event per hour as discussed earlier in this paragraph. Additionally, assume that only half of them remain and the other half is passers-by. All 50 users that remain are workers who did not leave the region in the interval 8.00a.m.-5.00p.m. The region has no residential homes. Therefore, in the 0-24h period, the sightings in the HPP database will be 100. Summing the sightings per hour of the DPUE database will result in: 50 + 50 + 9 = 500 sightings. Thus, the difference between DPUE’s and HPP’s sightings will be 400 = 50 * 8. This means that one sighting of each user that remained in the area is deleted; the number of events created by the users that remained in the area should have been 450 = 50 + 9 instead of 400. The effect of the difference is that the amount of ‘arriving’ users is underestimated and thus the destinations are also underestimated. This problem can be resolved with an iterative procedure where the result from the stepwise methodology described above is fed back to the first step by increasing the factor \( f_1 \) by the number of trips per destination calculated at step (3). Therefore, factor \( f_1 \) would become: \( f_1^k = DPUE_{region} - HPP_{region} + A_{region} \), where \( k \) is the number of iterations.

c) One note should be made for the difference of DPUE and DPU sightings. The DPU sightings actually show the active hours and not the events. If a user commuted to his place of work and was sighted at 8:00a.m. in the zone where his working place is located and then created one event per hour until he left to return home at 17:00p.m., he will have been sighted a total of 9 times. No matter how many more than 9 events the user created in that time interval, the total sightings will be equal to 9 because these are the 1-hour time intervals he was seen at. Therefore, the DPU dataset shows how many users were active in the 1-hour intervals. Consequently, this means that they are not equal to the number of events but are actually always less than the number of events. This is why this methodology requests a different dataset.

d) The temporal scale used should be a fraction of the working day preferably the same as the one used in chapter 4. The average day is created by multiple travel diaries on a sample of random individuals who compile their diary each for a random working day. In this case and due to the abundance of data, it is possible to average the trip matrices over a week. Summing and averaging over multiple weeks will damage the data because of the amount of seasonal variations.
5.2.2 A probabilistic approach for excluding ‘passers-by’

This method effectively aims to estimate the probability of counting an individual in a zone. Using the probability for every zone, the excess sightings (being multiple sightings) can be excluded while the probability of not using the phone and therefore not appearing in the dataset is also included. This method bears resemblance but is an adaptation of the method developed by Astarita & Florian (2001).

The authors elaborated on the use of mobile phones for Electronic Toll Collection – ETC. They believed that the mobile phones could be used to locate the user on the road and charge him/her. Their method aimed to provide an alternative to manual toll collection which increases travel time and operational costs of the highway. The authors had an impeccable understanding on the limitations of positioning a mobile phone at a time when 3G networks were not introduced which consequently means that phones had no internet connection capabilities. The problem which Astarita & Florian (2001) faced was to determine which vehicle can be identified in the system and which cannot. They hypothesized that the probability of detecting a vehicle that moves between an
Origin-Destination pair depends only on that pair and the interval that is used. As a result, the probability of detecting a vehicle moving from origin- \( O \) to destination- \( D \) in the time interval \( k \) is:

\[
(P_{OD})_k = \left( \frac{Nd_{OD}}{N_{OD}} \right)_k
\]  

where, \( N \) is the total amount of vehicles in the system and \( Nd \) is the number of detected vehicles in the system.

The problem that Astarita & Florian (2001) encountered is similar to the case of viewDat database. A mobile phone user can be sighted at any place between his origin \( i \) and destination \( j \). Assume that there are \( w \) regions between origin and destination. Then the probability of a user being a ‘passer-by’ is equal to the number of ‘passers-by’ over the total amount of users in the system, which are the ‘passers-by’ and the ‘arriving’.

What is most useful in the specific problem of the viewDat database is to determine the ‘passers-by’ at each region. But there are a lot of people crossing a specific region that come from different origins and go to different destinations. For example, Apeldoorn is crossed by A1 highway which is major pathway from Enschede and Germany towards Utrecht and the Randstad. Additionally, it is crossed by A50 which links Zwolle and other cities of Northern Netherlands with Arnhem and Nijmegen.

To determine the ‘passers-by’ of Apeldoorn, one must consider all possible OD pairs whose connecting routes cross Apeldoorn. The probability \( P_{ijw} \) of crossing Apeldoorn but not stopping is:

\[
P_{ijw} = \frac{PB_{ijw}}{T_{ijw}} \Rightarrow PB_{ijw} = P_{ijw} \cdot T_{ijw}
\]  

where, \( T_{ijw} \) is the amount of travellers between any origin and any destination crossing region \( w \), i.e., Apeldoorn, and \( PB_{ijw} \) is the amount of ‘Passers-By’ that started from any origin and are bound to any destination and whose route crosses Apeldoorn. This formula is equivalent to (23).

Both \( T_{ijw} \) and \( PB_{ijw} \) are unknown while formula (24) does not incorporate the probability of phone usage. Incorporating this probability is reasoned with an example illustrated in Figure 33. Consider 4 origin zones labelled as: Zero, 1, 2 and 3. Each region generates 100 trips towards the destination. Assuming that the probability of a person crossing and using his phone at any region is 50%, then half of the travellers from origin Zero will appear in zone 1. Consequently half of the travellers from zone 1 will appear in zone 2 along with half of the travellers from zone Zero who are also crossing zone 2. The probability of a person being located in zone 2 would be 50% * 100 + 50% * 100 = 100. Applying the same logic for zone 3 would result in 50% * 100 + 50% * 100 + 50% * 100 = 150 ‘passers-by’.

![Figure 33: An example of detecting passers-by](image)

Chapter 5 - Additional methods for extracting OD matrices from aggregated data
In the example above, all users are passers-by. Therefore, the probability of a user being a passer-by should have been equal to 100% and not 50%. It is because of the phone usage that the probability was reduced to 50% which in turn means that $P_{ijw}$ is the general probability of someone being a ‘passer-by’. What is not shown in the example is that the effects of phone usage are already applied and incorporated in the dataset; the actual sightings in the viewDat database are reduced because of phone usage. This means that in practice, instead of 100 trips being generated (not sightings) per region, the recorded values would be 50% reduced that the values shown in Figure 33 since it is assumed that the probability of locating someone in a region is 50%. Incorporating the phone’s usage probability $P_{usage}$ would make formula (24) equal to:

$$P_{usage} \cdot (P_{ijw} \cdot (P_{ijw} + T_{ijw})) = P_{usage} \cdot (P_{ijw} + T_{ijw})$$

$$p_{ijw}^{detected} = p_{ijw} \cdot S_{ijw}$$

where, $S_{ijw}$ is the number of sightings that is known from the HPP dataset or another aggregated GSM dataset of similar specification. The variable $P_{ijw}$ is unchanged. Comparing formulas (24) to (25) provides an important observation. It shows that the probability of phone usage is only useful for determining the total amount of users, those that are logged and those that are not sighted in the database because they did not create an event.

Probability $P_{ijw}$ is impossible to determine without direct measurements of users that are actually ‘passers-by’. The complement of this probability however is easier to determine because it is the probability of a person stopping in the area while travelling from the same origin. This notion is equivalent to determining the destination, mode and route. Probability $P_{ijw}$ is then: $P_{ijw} = 1 - P_{iw}^{arriving}$.

The computational load of determining such a probability is big but can be assisted with the use of modelling software. Determining the aforementioned probability is a standard operation in transport planning. This probability for every region can then serve as a factor while the application of formula (24) will lead to estimating the amount of ‘passers-by’.

The final probability for each region is a summation over all pairs of origins and destinations that cross the examined region which is:

$$p_{ijw}^{detected} = \sum_{i,j} p_{ijw} \cdot S_{ijw} \iff p_{ijw}^{detected} = \sum_{i,j} (1 - p_{iw}^{arriving}) \cdot S_{ijw}, \forall w$$

The methodology devised and discussed in this paragraph is a brute methodology that requires a lot of computational resources and may not be practical as other data-focused approaches. However, it is methodology that will result in estimating with good confidence the number of passers-by, more consistently than the previous methodologies discussed. It also can be extended for use in the application proposed by Astarita & Florian (2001) which is automatic toll collection.

### 5.3 Conclusions

The GSM problem can fundamentally be broken down to a single problem of excluding passers-by from the dataset. In essence, methodologies for determining whether a user participated in an activity or not, are required. Most studies use a spatiotemporal rule to determine when a user stopped to participate in an activity. These rules are not always justified. Nevertheless, Bayir et al. (2010) who did an extensive study on this matter note that any threshold above 30min will not have a great impact on whether the user appears as immobile or on the move.
Solutions for excluding the passers-by were proposed in this chapter. It is important to note that the reason for such methodologies to exist is in the case where a database already consists of multiple counts. These databases could be created from raw data that have no user identifier such as a hashed ID.

Most importantly, this chapter showed that extracting OD information can be done from simple databases. These solutions explore the opportunities of data fusion while using information from other databases, such as amount of workers in a region or counts of cars (in comparison to slow modes and train), may provide information besides origins and destinations. What can be concluded from this chapter is that GSM data can play a crucial and central role in transportation planning, but they can replace current methods only if they are enriched with more information.
6. Conclusions and Answers to the Research Questions

6.1 Answers to research questions

Answering the research questions is the main goal of this thesis work. The answer for the research questions can be found below.

- **What is the spatial and temporal aggregation used in existing transport planning models and in what level of transport analysis can the data be useful?**
  
  Existing strategic models at the national level, are aggregated spatially in zones that have the same population and comparable trip productions and attractions (Cambridge Systematics, Inc. and AECOM Consult, 2007). Temporally, strategic models are usually aggregated in three or four time periods which are morning and evening peak, rest of day and night time (after evening peak) (Fox, et al., 2003). Additionally, a number of models use peak shoulders to account for all the trips that are made during the peak hours (Chuck, 1999; Smit & Flikkema, 2010).

  The OD that was created from the viewDat database performed well in comparison to observed and modelled values. However, its power is diminished for kilometres below 7.5km as it was seen by the trip length distribution of Figure 26. This is the case because the spatial aggregation to BAG regions is coarse when zones contain cities. The OD is certainly useful only in the national level in spatial terms and not in a more detailed configuration. Additionally, because of the fact that only a morning peak matrix could be calculated it cannot be used in other temporal scales except, maybe, evening peak. Some strategic models use a mirrored matrix of the morning peak for the evening peak since they contain trips of the opposite direction than morning peak (home-to-work and work-to-home) (De Dios Ortúzar & Willumsen, 2011).

- **What are the limitations and inconsistencies of the data and how can they be solved?**
  
  A complete list of the limitations is given in 3.5. Some limitations are summarized below:
  - It has been found in chapter 3 that, months before and including June have to be excluded from further analysis for the reasons discussed in subchapter 3.6.
  - Categories of frequent, regular and incidental visitors are not connected to trip purpose. The use of these categories has to be justified for use in modelling.
  - There are 7 regions whose coding is duplicated with different data for each code such as Apeldoorn.
- There are data entries in the HPP dataset where origin and destination are the same but the amount of visitors is not 0.
- The multiple counts are damaging the information included in the OD matrix greatly.
- The size of the sample cannot be directly calculated. A reason why active users are different from residents should be given by Mezuro.

The following limitation and solutions have been found and proposed respectively.

Table 14: Cumulative table of the limitations and possible solutions

<table>
<thead>
<tr>
<th>Limitation</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Counts</td>
<td>Can be overcome by balancing origins towards destinations. It would be more accurate if Mezuro alters the specification of the HPP database.</td>
</tr>
<tr>
<td>Less than 15</td>
<td>Can be overcome by applying the methodology described in 4.3.2 for the viewDat database.</td>
</tr>
<tr>
<td>No correspondence of trips for zones that abstain less than 7.5km</td>
<td>The spatial aggregation should be altered to match the specification of traffic zones. Alternatively, the cells can be used as traffic zones.</td>
</tr>
<tr>
<td>Only morning peak matrix can be created</td>
<td>It would be more useful if the HPP database estimated “trips” for periods in the peak hours instead of the current periods.</td>
</tr>
<tr>
<td>Data cannot be disaggregated to periods of day (applies only to framing data)</td>
<td>It should be better to aggregate days per time periods instead for the whole days. Weekend data must be excluded.</td>
</tr>
<tr>
<td>Size of sample and phone activity</td>
<td>By not excluding the people that are not active enough, phone usage becomes a variable. It is better to only include the most active individuals as in most studies found in literature.</td>
</tr>
</tbody>
</table>

The issue of spatial aggregation which is the reason why there is no correspondence of trips for zones with distance less than 7.5 km, could be solved by creating new regions for the big cities and applying the centre-of-gravity approach used in this thesis and by Ratti et al. (2006). This is possible since the trips that appeared in the first kilometre bin are equal to the trips of kilometre bins of longer distance. The proposed factor, for distributing trips from an origin \( i \) to a destination \( w \) that is divided to \( j \) regions, is the following:

\[
T_{ij} = \frac{Workplaces_j}{Workplaces_w} \times T_{iw} \quad (27)
\]

Alternatively, instead of a proportionality factor, a spatial weight could be applied such as an exponential function similar to the one calculated in 4.3.2.

- If Origin-Destination matrices cannot be derived from the database, then what additional information is required from the data to achieve this purpose?

Since OD matrices were extracted from the database, this answer constitutes a general remark on improving the information enclosed in the viewDat database. The framing data resolve the most issues of the viewDat database such as multiple counts and spatial aggregation. Altering the spatial aggregation requires information about the accuracy of positioning users because it is possible that the coarse spatial aggregation for cities is a result of the inaccuracy of positioning the user. Nevertheless, the problem of multiple counts can be resolved easily by using the same concept as in the DPU data, the user is counted only where he/she made the majority of events.
An issue that needs to be addressed, and that the framing data do not solve, is the temporal aggregation. The amount of ‘excess trips’ and multiple counts create many problems to accurately estimate an OD matrix. A new period that is closer to the morning peak hour would be much more useful for that task. Therefore, instead of 10-12h period the 8-10h or 9-11h would be more convenient for estimating the OD matrix because these periods would contain trips congruent to the trip purpose usually seen in morning peak.

The type of trip should also be a required piece of information. The work location can be identified by a spatiotemporal rule or models available in literature such as spectral clustering described by Nurmi & Koolwaaij (2006). Finally, in order for the OD matrix to be equivalent to existing matrices, information about the mode used, should be included although it is an important topic of current research study.

What are the possible uses of the viewDat database besides deriving Origin Destination information?

One of the possible uses is the examination of the differences in working days. The LMS matrix and most other OD matrices are created for working days or, more specifically, for a representative working day. Willumsen (1982) showed that while the variations in flows might be small among day types of different weeks, the difference in the OD matrices cannot be neglected. However, most OD matrix studies and practical implementations do not account for the temporal variation because it would result in major changes in the trip generation and distribution models. Ortúzar & Willumsen refer to the temporal effects on the OD matrices as “inconvenient truth” (De Dios Ortúzar & Willumsen, 2011, p. 205). It would be cost inefficient to study changes within a month or a year especially when the main purpose of the matrix is to predict movements in 20 or 30 years. Given this prediction timeline, the difference in day types of different weeks, is not important. What is important and usually neglected is the difference in the base year. The OVIN study extends for multiple months and years. The temporal variations that are inherent in estimating the base year matrices are damaging the matrices and consequently, their predictive power.

The viewDat database can have validation uses for existing OD matrices. The estimation of the productions in most models is accurate because the underlying variables are measured systematically and consistently for many years while there has been in depth reviews of the estimation techniques. Additionally, the household surveys are linked with socio-economic information. GSM data and consequently the viewDat database cannot compete with the power of household surveys in estimating the productions. However, the attractions of models derived from household surveys are not as accurate as the productions. Most models would equilibrate the attractions towards the productions since the latter are deemed more accurate. However, the quality of spatial information is debatable. This is where the viewDat database can contribute the most, in calibrating/validating the attractions prior to the balancing towards the attractions or to validating the balanced attractions and to validating the values of trips for OD pairs that abstain more than 7.5 km.

Can Origin-Destination matrices be extracted from aggregated GSM datasets, how can this be achieved and how accurately do they represent reality?

OD matrices can be extracted from aggregated GSM data with the prerequisite that they contain information about where a user is mostly seen as the DPU dataset has and information about home location. More specifically, the solution requires the fusion of the information of at least two datasets one of which contains information about the origin. The dataset that does not contain origin can have even the simple form of events counter such as Erlang (see paragraph 5.2.1).

The method used in this thesis, uses the observation that trips in the end of peak hours are much less than the trips made within the peak. The data for 10:00 a.m. can be used as destinations. Caution must be put into the rounding of the values of the cells after the operation.
The matrix seems to be comparable to the Dutch national model, LMS and the observations made by the Centraal Bureau voor de Statistiek. The spatial aggregation of the datasets that this research work used, are not advised for strategic modelling because the trips made between zones that abstain less than 7.5 km are accumulated in the large zones of cities.

Additionally to the answers above, the following conclusions and observations specific to the viewDat database have been made in the course of the thesis:

- Understanding the dynamics of phone activity is paramount in order to scale the sample to the population and to the amount of phone users irrespective of whether they used their mobile apparatus.
- The division of type of visitors to frequent, regular or incidental cannot be related to purpose. Such an assumption will lead in ecological fallacy because it is known that there are more purposes than categories and that multiple purposes fit in one category.
- The sparseness of the framing data matrix (see Appendix C) is 84% while of the viewDat database is 99%. Nevertheless, both datasets are created using the same sample which raises the question of what information are excluded from the viewDat database. It could possibly be attributed to the spatial aggregation of the viewDat database which is more detailed in less populated areas which in general have a low number of trips.
- Scaling up the sample could be achieved the same way as takes place using the OViN data. The input is equivalent, travel diaries of individuals. It certainly cannot be achieved the same way as with the viewDat database because the data are not population counters. The best practice would be to use trip proportions of the synthetic LMS matrices for all modes provided that the synthetic matrices are available. Doing so would ensure that the same scaling up procedure that takes place in the OViN data is passed on to framing data.

6.2 Conclusions

6.2.1 Conclusions from the process of the research

Two important conclusions can be drawn for the SSIM index. First, the accuracy of the index is not adequate when the variance and/or the covariance is below 1. For that purpose, a slight alteration has been proposed in 4.6.2. Second, the use of the SSIM index can be extended to checking for sparseness. This can be achieved by transforming all the non-zero values of both matrices to 1. It was seen that simply comparing the percentage of cells that are zero of two matrices is not enough to indicate whether the sparseness in these matrices is of the similar nature. It is possible that completely different cells were zero in one matrix (meaning that different regions showed no trips between each other) than the other matrix or it could be that they were the same. The SSIM index could be used to account for spatial characteristics of sparseness.

In order to compare two matrices for sparseness, another problem was encountered: the spatial aggregation of the two matrices to be compared was different. In the case that one of the two matrices was made for a coarser aggregation than the other, the problem can easily be solved by using the tools of QGIS for aggregating to the coarser specification. However, the two specifications of the matrices that this thesis used, BAG and LMS regions, did not have such a straightforward relationship. There were regions in the BAG zoning system that were coarser than the LMS regions and vice versa. Therefore, it was decided to aggregate all the zones to a common reference system. This system was simply a map of rectangles that was placed over the map of the Netherlands. The regions of both systems were then aggregated to these rectangles. A simple rule was used to define the size of the rectangles: find the smallest amount of rectangles that contain...
only one centroid of LMS or one centroid from BAG. That ensured that most of the rectangles contained one centroid from LMS and BAG systems.

It should be mentioned that the process of scaling up requires the size of the sample. This has been noted before (Daas, et al., 2011) and was stated as one of the reasons not to investigate the data further. Most scientists decide to reduce the sample they use in order to avoid the aforementioned problem and in order to obtain only non-fragmented traces. Any database, aggregated or not, should include only a specified sample of most active users which is an acceptable trade-off that results in non-fragmented movements.

6.2.2 Extracting OD information from aggregated GSM datasets

The research conducted on the viewDat database shows that an OD matrix can be created from the data, that the ‘Less than 15’ problem can be resolved using transport modelling techniques and that the derived matrix can be implemented for calibrating other OD matrices or even direct use in modelling.

The main components for deriving OD information from aggregated GSM data, as they were seen with the viewDat database and beyond (see chapter 5), are the following:

- A database that contains either Erlang-type data or that contains information on where the user was seen the most.
- A database that contains information on direction. In the case of this thesis this database was the HPP database that contained information on the home locations. Nevertheless, another type of database, simpler than deriving home locations from a monthly dataset, could be used. This database should contain two pieces of information: (i) where the user is located from 1-5h (or any other period not close to the beginning of morning peak, notice that increasing the period increases uncertainty on home location but increases the accuracy of its estimation) and (ii) information on where these users are located at 10:00 a.m. if they are not at their home location.

In other words, the minimal requirements for extracting OD information from GSM data are the home locations and the locations were users have been more active.

This thesis shows that extracting origin destination information from aggregated GSM data is easier than it was originally thought. It is possible to derive aggregated information for use in aggregate methods of modelling which extends the arsenal of transportation engineers even further. Nevertheless, the term aggregated implies that some information is not retrievable due to the level of aggregation. Information such as purpose and mode cannot be retrieved without using either the raw data at the telecommunication operator’s side as Mezuro does, or using modelling techniques.

The user categories that Mezuro uses (frequent, regular and incidental) may be useful to study trip purpose but further insight into the dynamics of these groups is required. For example, if a user visits an area more than 10 times in a month which consequently means that it is extremely possible that he/she is a worker in that area, then why is the number of frequent visitors so high in the weekends? Additionally, why does the weekly variation of regular visitors (2-10 visits per month) resembles the variation of frequent visitors? And finally, why does the purpose of shopping resembles only the variation of incidental visitors when it is the purpose that most trips of a user are made for? There are simple rules that could assist in creating accurate and useful categories instead of only using the amount of times a person visited an area.
6.2.3 A comparison of different specifications of GSM data

The following SWOT analysis was compiled to analyse the aspects of the various GSM data specifications seen in this thesis with respect to the purpose of creating OD matrices.

Table 15: SWOT analysis on disaggregate GSM data

<table>
<thead>
<tr>
<th>Strengths:</th>
<th>Weaknesses:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Determine the trajectory of users</td>
<td>1. No connection to socio-economic data</td>
</tr>
<tr>
<td>2. Any level of spatial and temporal aggregation</td>
<td>2. Require analysis in order to be useful for transport planning applications</td>
</tr>
<tr>
<td>3. Size of sample can fit the demands of the study</td>
<td>3. Require adequate equipment to be handled (e.g., servers)</td>
</tr>
</tbody>
</table>

The viewDat data are in fact aggregated GSM data according to the definition given by Calabrese (2011). Their specification, limitations, strengths etc. are indicative of other aggregated GSM datasets. Therefore, they are used as a proxy for other commercially available GSM datasets.

Table 16: SWOT analysis on viewDat data

<table>
<thead>
<tr>
<th>Strengths:</th>
<th>Weaknesses:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Determine temporal and spatial interactions</td>
<td>1. Incosistencies in montlhy variations - No complete year of data</td>
</tr>
<tr>
<td>2. Determine visitation of area based on home location</td>
<td>2. Complicated or unclear specification and datasets’ definition</td>
</tr>
<tr>
<td>3. Massive amount information per hour of day</td>
<td>3. Spatial and temporal specification</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Opportunities:</th>
<th>Threats:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Derive OD information per day for every month the data are available for</td>
<td>1. The spatial and temporal specification may be unchangeable</td>
</tr>
<tr>
<td>2. Slight alterations of the specification can validate the assumptions</td>
<td>2. Missing data entries per month</td>
</tr>
<tr>
<td>3. Uses in dynamic modelling</td>
<td>3. Changes in the specification are exogenously imposed and not analysed</td>
</tr>
</tbody>
</table>
The so-called “Framing data” are a step away from aggregated GSM datasets and closer to disaggregate data. Although they still are aggregated GSM data, they are not similar to viewDat data. Furthermore, they fit the definition of aggregated GSM datasets given by Wolf et al. (2014b). They are therefore used as a proxy of another type of aggregated GSM data.

**Table 17: SWOT analysis on framing data**

<table>
<thead>
<tr>
<th>Strengths:</th>
<th>Weaknesses:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Origin-Destination information are directly available</td>
<td>1. The temporal aggregation is not helpful considering the purpose the dataset</td>
</tr>
<tr>
<td>2. Information on the users that participated in trips</td>
<td>2. Complicated or unclear specification and definition</td>
</tr>
<tr>
<td>3. Same spatial aggregation as LMS</td>
<td>3. The 'Less than 15' problem is not reversible</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Opportunities:</th>
<th>Threats:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Information about trip purpose could be included</td>
<td>1. The accuracy of individual's location estimation has to be analysed</td>
</tr>
<tr>
<td>2. OD matrices per period of day, i.e., morning -evening peak and rest-of-day</td>
<td>2. Changes in the specification are exogenously imposed and not analysed</td>
</tr>
<tr>
<td>3. Information about mode of transport</td>
<td></td>
</tr>
</tbody>
</table>

To the author’s opinion, improving the viewDat database should be of a bigger concern than improving the framing data. The reason is that the viewDat data use simple algorithms to determine the required information which do not insert uncertainty into the data. In essence the data are still raw. Framing data is an outcome of a model, a method called ‘framing’. The accuracy of the method needs to be established first in order to grasp the limitations of the dataset.
7. Recommendations and Future Research

7.1 Recommendations

This chapter will discuss two proposals, one for altering the dataset towards using it for calibration/validation of other matrices and one for extracting OD matrices. These two proposals could be similar because the OD matrix that is created from GSM data could be used for calibration. But, it is important to take into consideration that creating OD matrices is a computationally intensive process with complications. Therefore, the first proposal, about calibration, is given for the purpose of easing the complexity required to create OD matrices.

One important note concerns the size of the sample and the setup of the data collection procedure. The sample presently includes all mobile phones. But not all movements of mobile phones are included because phone usage varies based on circumstances such as purpose, mode and time of day. It is important to select the sample instead of using any available information that are of debatable accuracy. The best approach is to select a specific sample of most active and most mobile users. Both these statements have to be satisfied. It was discussed in subchapter 2.6 that younger generations will appear as more active. But it is possible that they do not appear as more mobile, i.e., that they remain in area which could be their city/home location. Therefore the sample should be carefully selected so that not only most active users are included but that these users that are selected also move in the spatial plane frequent enough to have equal likelihood to be seen in other regions as well as their home location.

7.1.1 Using the viewDat database for calibration/validation

In the case that the database is used for calibration/validation purposes, the sightings of the database should better resemble counts. Counts show the continuous flow $F$ of vehicles and persons although they are usually studied within a specified time interval or intervals $t$. The following holds for $k$ links (Cascetta, 1983):

$$A \ast t = F$$

(28)

where, $A$ is the assignment matrix derived from assignment models.
Chapter 7 - Recommendations and Future Research

The counts of the viewDat database have to resemble the flows of equation (28). Doing so requires the following alterations of the DPU and HPP datasets of the viewDat database:

- The HPP dataset should contain all available sightings in the same region without restricting the sightings to only one per user per region.
- The DPU dataset, or better the DPP, should be a database of events. This would make the dataset similar to the DPUE database discussed in paragraph 5.2.1

In this proposal, the database is considered to be altered to a specification similar to road counts. The added value of using GSM data instead of counts is threefold:

- The road counts require investment and funding to be placed, to be maintained and to extract the data. The GSM data do not require an investment to be placed while the maintaining process is negligible from the transportation engineering side since it is done by the telecommunication operator for another purpose which is to maintain the GSM network. This proposal should find many uses in countries where road counts are not well or extensively implemented.
- Road counts contain information about all cars. To determine if specific cars used specific exits, authorities use other systems such as plate-recognition software. This would not be necessary with GSM data because the information about specific users is known.
- Road counts refer only to road while GSM data can be used for counting other modes as well and especially suburban/rural trains.

Gathering information from mobile phones that are similar to counts requires that only specific antennas, or better yet, specific cells are analysed. By gathering information from cells that cover only roads and especially highways, the antennas could be used as counts. Therefore, instead of using all the cells in the Netherlands, only cells above highways and railways should be used. (Caceres, et al., 2007)

7.1.2 **Optimal specification of the viewDat database to extract OD matrices**

If the requested use of the dataset is for deriving origin destination information, then the proposed alterations concern mainly the HPP database. These were discussed in paragraph 4.6.1 but will be revisited here to provide an optimal (instead of an improved) specification. There are two important points to be made for deriving OD matrices. The first concerns the spatial aggregation and temporal aggregation of the database and the second the specification (or definition) of the databases.

The temporal and spatial aggregation should be changed with respect to one another. In other words, there should be a compromise as to how detailed each specification, temporal or spatial, can be. On the one hand, the specification should be such to avoid accumulation of trips in the first kilometre bins of regions that are in close proximity. On the other hand, a detailed spatial aggregation will allow for the ‘Less than 15’ problem to damage the database. Smaller regions generate fewer trips but if trips are measured over a day instead of, for example, an hour, then the ‘Less than 15’ issue might be extinguished.

The zones that are used have to be consistent to the original goal which is OD matrix creation. Thus, the characteristics of the zones cannot stray a lot from the characteristics of traffic zones. Therefore, the spatial aggregation should be consistent with any type of aggregation that is used in transportation modelling such as LMS, NRM or NVM (property of Goudappel Coffeng B.V.). The spatial aggregation that is chosen should dictate the detail of the temporal aggregation.

The temporal aggregation cannot vary much from the commonly used intervals in transportation planning. These are peak hours, rest of day with a possible distinction for night hours. The peak hours contain trips that start and end during peak hour. Due to peak spreading, there might be trips...
that begin within a peak hour but end after it and vice versa. For that purpose, the LMS uses peak shoulders. Therefore, the peak hours are not two 2-hour intervals in the morning and evening but are actually two 4-hour intervals with respect to the aforementioned constraint. The reason for including these time intervals in addition to the “rest of day” interval is that the trips made within these intervals are of similar purpose which is recurrent for working days. Recurring trips are those that have the same origin and destination. For example, business trips and shopping trips may have different destination within a week, month or even day. Home, education and work locations are much more stable and change rarely within a year.

It was seen with the framing data (see Appendix C) that if the LMS traffic zones are used with GSM data, then only the temporal aggregation of a month provided adequate results. This temporal aggregation is not useful. Even the temporal aggregation of a day may not be useful because there is no distinction of peak hours; all hours are included in a “complete day” matrix.

The temporal aggregation should be over peak hours for day types or for weeks. There is not a great deal of variation of visitors within days of the same week. Therefore, aggregating over days of the week should not be excluded.

The following are concluded for the specification in spatial and temporal terms:

- The spatial aggregation should be in line with traffic zones. Therefore the available spatial systems that can be used for strategic planning for the whole of the Netherlands should be (from more aggregated to less): LMS zones, NRM zones, NVM zones. It should be noted that cells of the GSM network should be created by the telecommunication operator based on network usage which is an indication of mobility. Thus the aggregation could simply be cells of the GSM network. These could be created using Voronoi tessellation (Li, et al., 2012; Tettamanti, et al., 2012) and aggregated to traffic zones using the centre-of-gravity method (Ratti, et al., 2006).

- The temporal aggregation should be in line with the transport planning models. Therefore peak hours should be used. Additionally, peak hours should include peak shoulders with respect to persons leaving from home for the morning peak and from work for evening peak, thus, excluding persons that participated in other trips than home-bound and work-bound. To achieve this, the work location must also be calculated. The trips could be aggregated for more than one day in order to avoid the “Less than 15” problem. For example, they could be aggregated for peak hours over multiple days or day types within a month.

All the above require that the home location (and work location) is known or else an OD matrix would be impossible to be created. For morning peak, the datasets should be created only for persons that travel from the home location. The destination can be derived using the recommendation of paragraph 4.6.1. For the evening peak, the work-location is required. This can be achieved using algorithms found in literature such as by Mastre et al. (2009) and Nurmi & Koolwaaij (2006).

In the case that both home and work location are known, the added value of GSM data greatly improves because, the home and work location can be connected based on the person who travelled. In other words, both home and work location would be known per user. This is not the case in classical modelling where home and work locations are calculated separately and are not connected per user (therefore, trip balancing methods are applied). Additionally, the attraction part of an OD matrix is usually poorly calculated (De Dios Ortúzar & Willumsen, 2011). This can change with the use of GSM data.

The rest-of-day matrix would be harder to create because of the trips that do not start from either home or work. The estimation of such a matrix requires an algorithm similar to framing. Aggregating temporally could take place for random hours between the peak hours or after evening peak. But, there could a shortcut to calculating such a matrix which requires the utilization of the work location.
Trips originating/ending in the work location within the peak hours have one known origin/destination. The counterpart of the origin/destination can be found by determining where the user created the most-events as it currently takes place in the DPU dataset (and is it was assumed for creating an OD matrix in chapter 4).

### 7.2 A recommendation for an improved methodology for applying the SSIM

The SSIM should not be used if the cells that are included in each of the frames used to calculate the SSIM are not spatially related. Wang et al. (2004) used the idea that pixels in close proximity show dependencies; thus, the pixels in a frame result in inter-dependencies. This is de facto true for images but is not for OD matrices or any other type of matrix containing spatial information. It depends on the arrangement of the matrix which usually depends on the coding of the regions. Due to the one-dimensional ordering capability of the regions’ coding, some additional spatial information is needed in order to apply the index.

The first specification proposed requires that the regions inside the frame are neighbours. Nevertheless, if one includes all the neighbours of the regions inside the frame, one would logically conclude that all regions of the matrix should be used. Therefore, each frame should include one examined region and its neighbours. Additionally, the frames used cannot be predetermined but have to change per iteration. Keeping the idea that frames move per row and columns as described in this thesis, the frames should include regions that are in proximity to an examined region or multiple which are not also necessarily, in close proximity in the matrix. Therefore, instead of a frame moving per column and per row, an efficient algorithm should move per examined region. The following algorithm is proposed (for uses besides OD matrices, it should be noted that rows will be referred to as origins and the columns as destinations):

```
Algorithm: Calculate frames for the MSSIM index for matrices

Input: OD$_1$ matrix, OD$_2$ matrix
Output: MSSIM
1: SET $i = 0$ and $j = 0$
2: for $i = 1, 2, ..., size(OD$_1$) //loop over origins/rows
3: for $j = 1, 2, ..., size(OD$_1$) //loop over destinations/columns
4:     $i = i + 1$ and $j = j + 1$
5:     FIND neighbours of the origin $i$ and STORE them as $O_i$
6:     FIND neighbours of the destination $j$ and STORE them as $D_j$
7:     $A_k = O_i \cap D_j$
8:     if $A_k \neq NULL$
9:         then $[A_k(Origin, Destination)]_1 = OD$_1$(Origin, Destination)
10:            $[A_k(Origin, Destination)]_2 = OD$_2$(Origin, Destination)
11:       Calculate SSIM for $[A_k]_1$ and $[A_k]_2$
12:     else continue
13:    end if
14: end for
15: end for
16: RETURN MSSIM
```

This process ensures that the frame for which the SSIM is calculated always contains regions that are in close proximity. From experience from applying this algorithm, statement (7) results in a very
small matrix or in a null matrix. This could be avoided by applying a “loose proximity” measure such that statements (5) and (6) also include regions that neighbours to the direct neighbours of the examined regions.

Finally, there could be other methods to select the regions included in matrix $A$ of the algorithm above such as:

- Select the regions by distance from the examined region for example by setting a threshold and including all regions that have a distance smaller than the threshold
- Instead of a distance, create a threshold for accessibility of regions and include only those that are above the threshold (higher accessibility means more accessible)

Because this thesis applied this measure and created conclusions based on the outcomes of it, no conclusions on the accuracy of the aforementioned proposal can be drawn.

### 7.3 Future research

#### 7.3.1 On the viewDat database and Mezuro data

State-of-the-art research with GSM data focuses on creating efficient algorithms to accurately determine the movements of individuals. It is important to determine and improve if necessary, the accuracy of positioning devices. The spatial aggregation to BAG regions was probably selected because of the overlapping coverage area of the antennas. This known issue is usually solved with the use of Voronoi tessellation \citep{Tettamanti2012, DeJonge2012, Nanni2013}. The regions created by Voronoi tessellation and that constitute the Voronoi diagram can be used as traffic zones or converted to traffic zones \citep{Li2012, Tettamanti2012}.

ViewDat database should find many uses in determining seasonal variations of data. That richness of information about every day movement was never before available. Additionally, the form of the data resembles more road count data than origin destination data. The ViewDat database can be applied for validating existing matrices. But the calibration procedure is an entirely different and big topic that has to be analysed under the specific scope. The issue is that the data of the ViewDat database are aggregated for all modes. It is also known that the LMS calibrates both car and freight simultaneously using Passenger Car Equivalent-PCE values. Therefore, it might be possible to calibrate matrices containing multiple modes such as the synthetic matrices of the LMS. This would also make the synthetic matrices useful for use at the base year level when now they are not.

The GSM data lack connection to socioeconomic data. But socioeconomic data would be of no relevance if multiple complete OD matrices can be extracted per day or even per hour. Growth factors could be estimated by simply comparing the OD matrices in matters of months and years. One important aspect to achieving such a goal is determination of mode. It would be useful if (revealed) mode choice could be derived from GSM data. Data fusion has equal or greater potential. OV-chipkaart data provide information about public transport. Thus the only mode distinction that needs to take place within the GSM dataset is among slow and fast modes, former being walk and bicycle and latter being car, train, bus, metro. This simplifies the required algorithm for mode distinction to a great extent but data fusion of GSM and OV chipkaart data is an ordeal on its own.

The spatio-temporal characteristics of individuals’ movements can prove useful for determining purpose from the data. The motifs that Smoreda et al. \citeyear{Smoreda2013} found can provide the practical background for such an endeavour. Alternatively, information on how long a user remained in a location, how many times he visited the location and its distance from other known locations, should be helpful in determining the purpose of the trip.
Finally, the addition of the work location should be a required piece of information such as home location. Algorithms in current literature are possible to be used to achieve this goal (Maestre, et al., 2009).

### 7.3.2 On aggregated GSM datasets

The subject of the characteristics of phone usage is recurring in this thesis. The connection of phone activity to travel purpose, mode etc. is paramount, e.g., to determine the size of sample. This thesis found that excluding this variable by selecting a sample of adequately active individuals is preferable. Nevertheless, understanding the movements of people using GSM data requires knowledge of underlying variables.

Understanding the characteristics of phone usage is extremely important. For example, if a person uses two phones for different purposes, then these phones have to be connected or their movements should be merged. De Mulder et al. (2008) who looked into the safety of location data from the computer engineer’s perspective hypothesized that even so, determining the movements of a person is possible. The authors however, failed to determine how it is possible. Additionally, phone usage could be connected to mode and type of activity, for example people should use their phone less when in a leisure activity (see paragraph 0). Phone usage will give a better insight in the person’s activity and movements.

It is also hypothesized in this thesis, that aggregated GSM data could be very useful as an equivalent to counts. Especially in countries where road counts are not already installed and their instalment might be economically unviable, aggregated GSM data can prove a valid alternative. For that purpose, it would be interesting to determine the possibility of using selected antennas on the side of the highways or in locations that they could resemble the functionality of a road count. It could also be a politically correct alternative to maintain the functionality of the dataset against worries on privacy, since road counts also measure movement information but do not intrude privacy.

Another important subject is the creation of spatiotemporal rules for determining other locations besides home. Since GSM data cannot be related to socioeconomic data it could be more useful to use empirical rules for determining locations of leisure, work and education. Doing so requires detailed information on revealed trip purpose which could be derived from the OVIN study. It would also be interesting for connecting the work of Smoreda et al. (2013) to tours. The work of the authors can assist in creating more accurate tour-based models and to estimate the inaccuracy of trip based models who only account for one trip of the traveller.

Finally, the SSIM index that was used proved to be a powerful and useful tool. Concerning the methodology for its application, the size of the frame has to be analysed. This thesis used a 10x10 matrix because there is no evidence to support why any other size would be better or worse. Each frame has to contain zones that are related. Zones that are related spatially are usually neighbours. Therefore, a variable frame where only zones that are neighbours to a specified zone are included may be a better idea. On the other hand, the spatial relationship of zones can extend beyond neighbours. In any case, more research into the appropriate size of the frame should be performed.
The interest of the scientific community to create OD matrices that could replace household surveys is justified by the fact that current sampling and modelling techniques are trending towards alternative and more efficient methodologies. But, this is not the only use of GSM data. Aggregated GSM data seem to be able to provide OD information given a strict but simple specification which is a step towards efficiency. However, the uses of aggregated GSM data could be on replacing road counts or on studying temporal variations of the existing OD matrices, aspects that it was not possible to study before.

In the beginning of this report, a figure was given showing the interest on Big Data as it is portrayed in the number of searches at Google search engine. The following figure shows the interest in google searches for the term “Location History”.

![Location History Index](http://www.google.com/trends/explore#q=location%20history)

The term was chosen based on a current emergence of the fear of people being watched by the corporate giant Google. The company stored the information of where a user was logging into its services but did not mention how and where it will use this information. The locations were stored online and available only to the individual through a password sign-in. Nevertheless, the information
was gathered on an 'opt-out' basis, which means that the user was sending the information until he/she specified otherwise. Although the curve does not show that the current research on GSM data is threatened, it shows that users are becoming aware on the information they transmit.

In the case that the public sentiment is strong enough, this could result in shutdown of companies and on adoption of laws to protect information privacy. As an example, InBloom, a (big) data analytics company, was forced to shut down after the cancellation of 6 out of its 9 contracts with states of U.S.A. The states that cancelled the contracts claimed that it was because of privacy concerns of the civilians (Kharif, 2014). A number of other companies faced lawsuits or shutdown due to privacy concerns such as Nordstrom and Facebook. Therefore, would it be wiser to delve more into the personal information or be able to achieve the goal, which is efficient transportation, using as much less data as possible? This thesis has successfully engaged the latter.


[Accessed 3 October 2014].


Lessons Learned – A short reflection on the research process


De Dios Ortúzar, J. & Willumsen, L. G., 2011. Modelling Transport. 4th ed. Southern Gate Chichester, West Sussex, United Kingdom: John Wiley & Sons Ltd.


Lessons Learned – A short reflection on the research process


Lessons Learned – A short reflection on the research process


Kwan, M., 2012. Visualization and analysis of mobile phone location data, Melbourne, Australia: Royal Melbourne Institute of Technology (RMIT University).


Patel, S. N. et al., 2006. Farther Than You May Think: An Empirical Investigation of the Proximity of Users to Their Mobile Phones. 8th International Conference, UbiComp 2006 Orange County, CA, USA (pp. 123-140), Springer Berlin Heidelberg.


Lessons Learned – A short reflection on the research process


Lessons Learned – A short reflection on the research process

Personally I believe that I have a structured thought process. I cannot leave a subject unfinished before moving to the next one which resulted in spending a great amount of time in subjects that were not of such worth to the overall picture. What I have learned from this, which was assisted by the regular meetings with my supervisors and the assessment committee, is that in any project it is imperative to take steps back and take a look at the whole picture.

Although I knew where each step was going, I was not able to connect every step to the overall picture without finishing it first. That is why I believe that always asking “why am I doing this” and “how this is helping in answering the main research question” are questions that should always be answered in any step of any research.

Another important lesson that I received is that more information does not mean better understanding. This report includes only a portion, even if it is the biggest, of the work I put into the subject. I learned that a report should include the most relevant work and not all the work. Therefore, executing the research in a way that only relevant information are gathered requires experience that I did not have but that I received by finishing it. And of course, struggling for finding the essential information must be in alignment with the thought that “more is always better than less”.

Finally, I had many difficulties but hope to have managed writing a concise and to-the-point report. I started writing the report from the early stages of the research. I thought that in that way, I will not have to struggle remembering the structure and the findings of each part of the research. That did help me in following the course of the research better than without doing it but, it also resulted in having many parts of the report written in different times. Thus, when I was correcting the report, I kept forgetting what I have written and had to redraw the whole report in order to align all these parts with the structure of the chapters, subchapters and research questions. The advice to me and others is to not start writing the report too soon or too late which insinuates that there is a decision point. Instead document the findings in notes that are easily retrievable and are categorized based on topic instead of time. That way one would be able to follow the process of the research without reading the main text of the report.

For performing this research work, I had to move from Delft and live in another city of the Netherlands, Deventer, where the headquarters of Dat.Mobility B.V. are located. I now have a reference point for any other professional environment that I will put myself in. I also have gained
immense experience from adjusting in another city. For example, I had to redefine my living habits since I had to live in a shared accommodation, I had to make new social circles since all my friends were far away, I had to operate in a mostly all-Dutch environment (in comparison to the international environment of Delft) which required that I learn enough Dutch for at least my everyday transactions and all that while I had to adjust to a working environment. I certainly feel blissful for having this opportunity and would recommend it to any of my fellow students. The effort is certainly worth the gain in personal and professional terms.