

MASTER THESIS

The Attractiveness of Cities

Determining city attractiveness using mobile phone location data

Author:

Jens VAN LANGEN

Student number: 4203879

Master Business Informatics

Supervisor:

Dr. Sietse OVERBEEK

(Utrecht University)

Second Supervisor:

Dr. Floris BEX

(Utrecht University)

External Supervisor:

Wim STEENBAKKERS

(Mezuro)

Second External Supervisor:

Menno DE PATER

(Decisio)



Universiteit Utrecht



mezuro



DECISIO

Faculty of Science

Department of Information and Computing Sciences

March 2016

Declaration of Authorship

The chapters 5 and 6 of this thesis are the result of collaborative research performed by fellow graduating student Johan MEPPÉLINK and me, Jens VAN LANGEN. The research presented in these two chapters formed a fundamental part of both of our research projects. Therefore, the research has been performed in collaboration but the writing of the initial version of the chapters has been divided. The chapter named '[Mobile Phone Location Data as a Data Source](#)' was originally written by me, Jens VAN LANGEN. Johan MEPPÉLINK adjusted the chapter to fit it to his research. Conversely, the chapter named '[Trip Motive Prediction](#)' was originally written by Johan MEPPÉLINK and has been adjusted by me, Jens VAN LANGEN, to fit into the context of my research. Finally, the appendix named '[Cell Radius Analysis](#)', appendix F, was also written by Johan MEPPÉLINK and has been adjusted by me, Jens VAN LANGEN, to fit it into the context of my research. The other chapters and content presented in this thesis are the product of my own work.

UTRECHT UNIVERSITY

Abstract

Faculty of Science

Department of Information and Computing Sciences

The Attractiveness of Cities

by Jens VAN LANGEN

The benefit of city visitors, i.e. people who live outside the city and come to visit, is that they positively affect income and employments levels in the city (van der Borg, Costa, & Gotti, 1996). Therefore, it is important for cities to develop a coherent set of goals and strategies directed towards attracting specific groups of people, rather than to promote at random (Petrişor-Mateuţ, Orboi, & Popa, 2013). To attain these goals cities can mobilise assets that can producing changes in the attraction and/or retention of specific segments of the population (Servillo, Atkinson, & Russo, 2011). However, to support their decisions to mobilise a certain asset, policy makers need information on how certain policy introductions or changes in the built environment, such as the placement of a new mall, will influence the attraction and/or retention of specific segments of the population.

This study shows a promising and novel approach to use mobile phone location data as a data source to empirically determine the relative importance of factors that influence city attractiveness. The method can be used for a variety of research fields including city centre retail attractiveness, urban planning and tourism destination attractiveness to determine the relative importance of factors that influence city attractiveness. The results obtained by applying our method seem logical but cannot be validated using existing literature, because our application was too broad to be able to compare it to earlier studies. Furthermore, it is very important to understand that the attractiveness of a city differs for various groups within our society (Sinkienė & Kromalcas, 2010). Therefore, this study has developed and implemented a model that can predict the most likely trip motive based on a set of trip attributes such as the arrival time and the day of week. Additionally, this study has proposed an improved method to extrapolate the sample, that is present in the mobile phone location data, to the travelling population. The main advantage of the new method is that it takes into account demographic differences of the population, which results in a more accurate representation of the actual population. The results are evaluated to the road-side measurements data by (Meppelink, 2016) and show a correlation ranging from .92 up to .98, depending on the situation.

Acknowledgements

I would like to thank everyone who supported me during my research. In particular my supervisors who have supported me through difficult times and provided plenty of good quality feedback: Dr. Sietse Overbeek, Dr. Floris Bex, Drs. Wim Steenbakkers, and Menno de Pater MSc. Next to that I would like to thank my colleagues at Mezuro and Decisio for their friendliness, laughs and very extensive lunches. Thanks to Johan Meppelink, my friend, classmate and colleague for the collaboration during the research. Thanks to Flaticon.com for letting me use their icons. Last but not least I would like to thank the people who have contributed in this research by giving interviews, providing data or have helped in any other way. Thank you all.

Contents

Declaration of Authorship	i
Abstract	ii
Acknowledgements	iii
List of Figures	vii
List of Tables	ix
Acronyms	xi
1 Introduction	1
1.1 Problem Statement	2
1.2 Report Outline	3
2 Research Questions	4
2.1 Research Questions	4
2.2 Research Scope	6
3 Research Approach	7
3.1 Overall Approach	7
3.2 Literature Review	10
3.3 Data Quality Assessment	11
3.4 Trip Motive Development	11
3.5 Data Analysis	12
3.6 Tools	13
3.7 Cooperation with Mezuro and Decisio	14
4 Related Literature	15
4.1 Introduction	15
4.2 Urban Competitiveness	16
4.3 City Centre Retail Attractiveness	21
4.4 Tourism	22
4.5 Urban Planning	24

4.6	Conclusions	25
5	Mobile Phone Location Data as a Data Source	27
5.1	Introduction	28
5.2	Obtaining Location Data from Mobile Phones	28
5.3	General Data Characteristics	34
5.4	Limitations	35
5.5	Scaling Devices to the Travelling Population	40
5.6	Improving the Scaling Method	44
5.7	Evaluating the Improved Scaling Method	50
5.8	Conclusions	52
6	Trip Motive Prediction	53
6.1	Business Understanding	53
6.2	Data Understanding	54
6.3	Data Preparation	58
6.4	Modelling	61
6.5	Evaluation	66
6.6	Implementation	72
7	Data Analysis	75
7.1	Introduction	75
7.2	City Selection	76
7.3	Data Collection	78
7.4	City Attractiveness	81
7.5	Data Preparation	83
7.6	Conclusions	84
8	Discussion & Results	86
8.1	Principal Component Analysis	86
8.2	Principal Components	91
8.3	Determining City Attractiveness	96
8.4	Relative Importance of Factors Influencing City Attractiveness	97
8.5	Relative Importance per Trip Motive	100
8.6	Discussion	101
8.7	Proposed Method	102
9	Conclusions	104
9.1	Introduction	104
9.2	Findings	105
9.3	Limitations	106
9.4	Future Research	107
	References	109
A	Interview on City Attractiveness	117

B Proposed Method	121
C Motive Prediction in OViN using R	125
D Factors influencing City Attractiveness	131
E Urban Competitiveness Model	133
F Cell Radius Analysis	135
G Developing the Improved Scaling Factor	138
H Regression Results	142
I Attribute and Data Source Mapping	146
J Descriptive Statistics	151
K Relative importance per trip motive	155
L City Area Selection	157

List of Figures

3.1	crisp-dm	8
3.2	A schematic overview of the chapters in this thesis, the Cross Industry Standard Process for Data Mining (CRISP-DM) phase to which each chapter is mapped and (if applicable) the final deliverable of the chapters.	9
4.1	The first part of the proposed method is applied in this chapter. The other activities in the method are collapsed as these will be applied in chapter 7 and 8.	16
4.2	Conceptual urban competitiveness model (Sinkienė, 2008).	18
4.3	Urban competitiveness maze (Begg, 1999).	20
4.4	The four-step travel-demand forecasting process (Martin, McGuckin, McGuckin, & McGuckin, 1998).	24
5.1	From left to right: a cell tower with antennas for various frequencies (2G/3G/4G), a camouflaged cell tower, and a cell tower located on top of a building.	29
5.2	The adjustable parameters of a single cell located on a cell tower.	30
5.3	The cells that are connected to when travelling the red dotted line. A destination is defined as a location where a user stays for 30 minutes or more and is depicted as a star. The cells are obtained from the mobile phone data, while the route in red is obtained from Global Positioning System (GPS) data.	32
5.4	The average number of events per user in 2015.	34
5.5	Data Collection	37
5.6	Comparison between Onderzoek Verplaatsingen in Nederland (OVIN) and the mobile phone location data of the relative number of trips per travel distance. Both datasets have their number of trips set to 100% at a travel distance of 26 kilometres.	39
5.7	Average scaling factor per Meuzuro area over the first nine months of 2015.	42
5.8	Geographical variation of the scaling factor corrected for the travelling population.	43
5.9	The penetration of subscribers per age group per gender (Offermans, Priem, & Tennekes, 2013).	45
5.10	The chance of making a trip greater than 10 kilometres per age group. This data is extracted from OViN.	46
5.11	The improved method to determine the scaling factor, which can be used to extrapolate the sample to the travelling population.	47
5.12	This inductive loop detector system is ubiquitous in the Dutch road network. The vehicle counts obtained through this system can be used to evaluate the travelling population inferred from the mobile phone location data.	51
6.1	An example of a Probability Estimation Tree (PET).	63
6.2	An example of a PET.	65

6.3	This figure depicts the axes of evaluation. Potential models will be tested on three levels of granularity and tested using a variety of settings for minleaf and m-estimation weights.	68
6.4	This figure depicts the relative importance of each attribute that is used in the final model to predict the trip motive. The total variable importance is scaled to 100%.	73
6.5	The comparison of the distribution of trip motives between OViN and Mezuro aggregated on the national level.	74
6.6	The comparison of the distribution of trip motives between OViN and Mezuro during morning rush hour, i.e. 6am-8am.	74
7.1	The second part of the proposed method is applied in this chapter. The other activities in the method are collapsed as these are applied in the chapters 4 and 8.	76
7.2	This figure depicts the impact of the minimum-of-15-rule. The data is based on trips from municipalities to Mezuro areas and is the average per day over the course of a month. The trips omitted is the difference between this level and the province to Mezuro area level. Hence, on average, the more visits a Mezuro area has, the less trips are omitted.	77
7.3	This figure depicts how the supply of alternatives of the periphery has been determined. The supply of alternatives is calculated by subtracting the number of amenities within the biggest radius from the number of amenities within the smallest radius.	81
7.4	The effects of the trips performed by a person who lives in the orange-coloured city on the attractiveness for each of the cities involved.	82
8.1	The final part of the proposed method is applied in this chapter. The other activities in the method are collapsed as these were applied in earlier chapters 4 and 7.	87
8.2	The number of principal components to retain is determined based on the minimum Minimum Average Partial (MAP) score. The results indicate that the optimum number of principal components to retain is five.	88
8.3	An overview of the scores per city per Principal Component (PC).	92
8.4	Scatterplots between the five PCs and the measures of city attractiveness.	98
8.5	Relative importance of the five PCs per trip motive per measure of attractiveness.	100
8.6	The proposed method to determine the relative importance of factors that influence city attractiveness.	103
B.1	The proposed method to determine the relative importance of factors that influence city attractiveness.	122
E.1	Urban competitiveness model (Sinkienė, 2008)	134
F.1	The percentage of trips observed in the GPS trace that can also be found in the mobile phone location data.	136
F.2	Percentages of origins correct for all four algorithms.	137
G.1	The improved method to determine the scaling factor, which can be used to extrapolate the sample to the population	139

List of Tables

3.1	A sample of the OViN data.	12
5.1	A part of the cell plan table, which contains the parameters of every cell.	30
5.2	An example of the Call Detail Records (CDRs) that are processed to obtain mobile phone location data.	31
5.3	The number of events handled by cells up to a specific radius.	33
5.4	A simplified and fictional example of the calculation of the Origin-Destination (OD) scaling factor for a single Mezero area for a single day. This example only includes the dimension age group (grouped by 20 years). The full method also includes the distinction between Saturdays, Sundays, normal workdays (Monday to Friday) and workdays during holidays but these are omitted in this example to the improve readability.	48
6.1	Trip attributes present in the mobile phone location data.	57
6.2	Trip attributes used in modelling and evaluation.	60
6.3	Results on country level. Showing how often the model's predictions are indistinguishable from the true distribution of trip motives. In total 20 Chi-square tests per field are performed.	69
6.4	Results on province level. Showing how often the model's predictions are indistinguishable from the true distribution of trip motives. In total 12 Chi-square tests per field are performed, which is equal to the number of provinces in the Netherlands.	70
6.5	Results on municipality level. Showing how often the model's predictions are indistinguishable from the true distribution of trip motives. In total 25 Chi-square tests per field are performed, which is equal to the number of selected municipalities.	70
7.1	The Dutch cities included in this analysis.	78
7.2	These data sources and tables are used in mapping the factors that influence the attractiveness of a destination to data to represent those factors.	85
8.1	The results of the Principal Component Analysis (PCA). For each variable it is noted how it is loaded onto each of the five PCs. Loadings greater than .6 are in bold. The communality is the amount of shared variance. The uniqueness is a measure of the unique variance explained by the variable. The complexity is the Hofmann's index of complexity, which indicates on average how many components are used to explain each variable (Hofmann, 1978).	88
8.2	The attractiveness of each city per measure of city attractiveness.	96
8.3	Regression Results	98
8.4	The relative importance of the PCs.	99

B.1	Descriptions of the activities and sub-activities that are used in the method depicted in figure B.1	123
B.2	Descriptions of the concepts that are used in the method depicted in figure B.1	124
D.1	Factors that influence the perceived destination image (Beerli & Martin, 2004)	132
G.1	Descriptions of the activities and sub-activities that are used in the method depicted in figure G.1	140
G.2	Descriptions of the concepts that are used in the method depicted in figure G.1	141
H.1	Regression results per trip motive for the absolute measure city attractiveness (A_j)	143
H.2	Regression results per trip motive for the measure city attractiveness per inhabitant (P_j)	144
H.3	Regression results per trip motive for the ratio between attracted and non-retained visits (R_j)	145
I.1	These data sources and tables are used in mapping the factors that influence the attractiveness of a destination to data to represent those factors.	147
I.2	This comprehensive list of factors from Beerli and Martin (2004) that influence destination attractiveness shows how data is mapped to each factor. For each factor the relevance of this factor for this research, the source and column the data is obtained from, how the factor is implemented and expressed in this research, is elaborated.	147
J.1	Colors to identify the literature source.	151
J.2	Descriptive statistics for the variables that characterise a city.	152
K.1	Relative importance of the components per trip motive for the absolute measure city attractiveness (A_j)	156
K.2	Relative importance of the components per trip motive for the measure city attractiveness per inhabitant (P_j)	156
K.3	Relative importance of the components per trip motive for the ratio between attracted visits and non-retained residents (R_j)	156

Acronyms

CBS	Centraal Bureau voor de Statistiek.
CDR	Call Detail Record.
CRISP-DM	Cross Industry Standard Process for Data Mining.
DM	Data Mining.
EU	European Union.
GPS	Global Positioning System.
HH	Household.
KD	Knowledge Discovery.
MAP	Minimum Average Partial.
MPN	Mobiliteitspanel Nederland.
MRQ	Main Research Question.
MS	Mobile Station.
NDW	Nationale Databank Wegverkeergegevens.
OD	Origin-Destination.
OVIN	Onderzoek Verplaatsingen in Nederland.
PC	Principal Component.
PCA	Principal Component Analysis.
PDD	Process Deliverable Diagram.
PET	Probability Estimation Tree.
SEMMA	Sample Explore Modify Model Assess.
SIM	Subscriber Identification Module.
SMS	Short Message Service.
SQL	Structured Query Language.
SRQ	Sub Research Question.

Chapter 1

Introduction

In our contemporary society over half of the world's population (54%) lives in urban areas and this number is expected to grow to 66% by 2050 ([United Nations, 2014](#)). Continuing population growth and the trend of urbanisation are projected to add 2.5 billion people by 2050 ([United Nations, 2014](#)), meaning cities are becoming increasingly popular places to live in. Most notably, the United Nations have projected that the population growth from 2015 to 2030 will be fully absorbed by cities, adding 1.1 billion urban dwellers to the existing population ([Cohen, 2015](#)).

Conceptually, cities are just high density agglomerations of firms and people living in residential areas ([Glaeser, 1998](#)) that function as centralities within a region for economic activity, innovation and employment ([Cohen, 2015](#)). Moreover, they also offer other advantages that are important for sustainable growth like reduced transport costs for basic services (such as water and electricity), goods, people and ideas ([Glaeser, 1998](#)).

Sustainable growth of cities is a policy dimension of the European Union (EU). In a study by [Servillo et al. \(2011\)](#) on the introduction of the concept of city attractiveness in EU policies they state *"... it is clear that diversity is a factor of attraction that can be utilised to generate growth by both attracting investment and mobile populations while retaining existing residents."* Using policies governments can mobilise assets that influence the attractiveness of cities. This attraction is defined as the ability to attract or retain specific target groups ([Servillo et al., 2011](#)).

To support their decisions, policy makers need data on how certain policy introductions or changes in the built environment, such as the placement of a new mall, will influence the attraction of visitors. Currently, the options to evaluate these effects are limited to local ad-hoc

solutions like visitor counts using wireless signals (e.g., Wi-Fi and Bluetooth) or using cameras (see the interview in Appendix A). Examples can be found in [Li, Zhang, Huang, and Tan \(2008\)](#); [Weppner and Lukowicz \(2013\)](#); [Xi et al. \(2014\)](#). The biggest downside to cameras, Wi-Fi and in particular Bluetooth is the limited range, which is of the order of 10 to 20 meters ([Sadhukhan, Chatterjee, Das, & Das, 2010](#)).

Another possibility is to consult one of the many rankings that exist and serve as a way to compare cities based on their level of attractiveness (for example [Global Power City Index, 2014](#); [A.T. Kearney's Global Cities Index, 2015](#)). However, many of these rankings are unreliable, subjective or fraudulent ([Petersen, 2005](#)) and their parameters cannot be adjusted. Therefore, the objective of this research is to assess how mobile phone location data can be used to determine the relative of factors that influence city attractiveness. For this purpose operator mobile phone location data, which is data collected by the operator each time a mobile device connects to a cell tower, form an interesting new data source. This type of data can give insight into how city attractiveness changes over time while providing unprecedented detail into users' travel behaviour.

1.1 Problem Statement

The benefit of city visitors, the people who live outside the city and come to visit, is that they positively affect income and employments levels in the city ([van der Borg et al., 1996](#)). Therefore it is important for cities to develop a coherent set of goals and strategies directed towards attracting specific groups of people, rather than to promote at random ([Petrișor-Mateuț et al., 2013](#)).

To the best of our knowledge no single study exists that uses mobile phone location data to develop an understanding of what the relative importance is of factors that influence city attractiveness. The factors that influence city attractiveness include, but not limited to aspects such as socio-economic capital (e.g., income levels), socio-cultural heritage (e.g., the amount of museums in the city), percentage of retail (e.g., amount of retail in the city), and demographics (e.g., average age of city dwellers).

Current academic literature does not use mobile phone location data to determine the relative importance of factors that influence city attractiveness, which is no surprise as this data source is quite new. However, there have been studies that confirm the feasibility of this idea. For

example, [Becker et al. \(2011\)](#) show that mobile phone location data can be used as a data source to group city dwellers and city visitors based on their mobile phone usage patterns. In this study we aim to go beyond this idea by predicting the most likely motive of a trip based on known trip and user attributes such as arrival time and whether a trip is to or from home. Scientific progress is based on high quality information ([Pampel et al., 2012](#)). Therefore, the goal of this study is to assess how mobile phone location can be used to determine the relative importance of factors that influence city attractiveness.

1.2 Report Outline

Chapter 2 provides all details on the research questions and the scope of this research. Chapter 3 delineates the approach and the methods that have been used to perform the actual research. Chapter 4 reports the findings of the literature review on the relevance of existing research on factors that influence city attractiveness. Chapter 5 serves as an introduction and a critical review of the use of mobile phone location data as a data source. At the end of this chapter we propose an improved method, that takes into account demographic differences in the possession of mobile phones and the likelihood of travelling, in order to scale the sample to the travelling population more accurately.

Chapter 6 describes the development and implementation of a model to determine the most likely trip motive based on a set of trip attributes, such as the trip distance, time of departure and day of week. Chapter 7 elaborates on the on the data that is used to represent the factors that influence city attractiveness. Next to that, the concept of city attractiveness is operationalised and the cities are selected which are included in the analysis. Chapter 8 presents the the relative importance of factors that influence city attractiveness using mobile phone location data and discusses the results. Additionally, a method is proposed to obtain the relative importance of factors that influence city attractiveness using mobile phone location data. Chapter 9 provides answers to the research question and provides directions for further research.

Chapter 2

Research Questions

This chapter provides all details on the research questions and the scope of this research. Section [2.1](#) introduces the main and sub research questions of this research, while section [2.2](#) describes how this research is scoped. The approach that was used to answer these research questions is discussed in the next chapter, i.e. chapter [3](#).

2.1 Research Questions

The goal of this study is to assess how mobile phone location data can be used to determine the relative importance of factors that influence city attractiveness, which is stated in the Main Research Question (MRQ):

Main Research Question

How can operator mobile phone location data be used to determine the relative importance of factors that influence city attractiveness?

Within the context of this research city attractiveness is defined as *the ability of a city to attract visitors and retain residents*, which is based on the definition of city attractiveness by [Servillo et al. \(2011\)](#) who defined city attractiveness as 'the ability to attract or retain specific target groups'.

To better handle the size of the research, the [MRQ](#) is split into several Sub Research Questions (SRQs), which will be discussed next.

In order to determine the relative importance of factors that influence city attractiveness it must first be determined which factors influence city attractiveness. Therefore, the following SRQ is stated:

SRQ1: What is known in literature about the factors that influence city attractiveness?

A popular saying is "garbage in, garbage out", which implies that the quality of the output is dependent upon the quality of the input (Lidwell, Holden, & Butler, 2010). Hence, it is essential to assess the quality of the data that is being used to determine a city's attractiveness, i.e. the mobile phone location data, in order to obtain valid results. Consequently, the following SRQ is posed:

SRQ2: What is the quality of the mobile phone location data?

Using mobile phone location data it is possible to determine the number of in- and outbound trips of a city. This serves as a base when operationalising the concept of city attractiveness later in this report. However, it is very important to understand that the attractiveness of a city differs for various target groups (Sinkienė & Kromalcas, 2010). To derive information about why some cities are more attractive than others it must be determined why people come to a city. Therefore, the destination choice (and the relative importance of the factors that influence this choice) is dependent upon the trip motive (Levinson & Kumar, 1995). For example, people who are interested in shopping are more likely to base their choice for visiting a city on the offering of retail as opposed to the offering of hotels. The reason for performing a trip is what is defined as the trip motive in this research.

Hence, since the relative importance of factors that influence city attractiveness can differ per trip motive it is necessary to determine the motive of a trip, which is stated by the following SRQ:

SRQ3: How can trip motives be determined from mobile phone location data?

Answers to the previous three research questions are a prerequisite for providing a theoretically informed answer to the MRQ of this research. The scope of this study was limited in several ways, which is discussed in the next section.

2.2 Research Scope

This section delineates three ways in which this research is scoped. The first limitation regarding the scope of this research is the applicability to medium-sized and large cities. The reason for this is two-fold. Firstly, from a data perspective it can be argued that the mobile phone location data is more accurate for larger cities. This is due to privacy legislation that does not allow Mezero to retrieve results that contain less than 15 users heading from a specific origin to a specific destination. Secondly, the lowest level of granularity at which the trips can be represented is at the level of a Mezero area. Mezero has split the Netherlands into 1,259 different areas. The larger cities consist of multiple Mezero areas, to better facilitate analysis, while the smaller cities and towns are combined with the urban and rural areas surrounding it. They did this because there is a trade-off between the accuracy of the analyses and the time it takes to produce usable data from the [CDRs](#) that is ready to be analysed.

The second limitation is regarding type of visit. The attractiveness of cities can be viewed from the perspective of many different groups (e.g., city dwellers, shoppers, school children, Sunday visitors, inner city visitors) ([Jansen-Verbeke, 1988](#)). However, in this study only the attractiveness from visitors who live outside of the city and come to visit the city *temporarily* is taken into account. Hence, studying permanent migration, i.e. moving to a new residence, is not the goal of this study. There are two reasons for this. First of all, attractiveness for visitors is more interesting because city visitors spend money in the city generating direct and indirect turnover for all types of businesses resulting in employment rates and a more vibrant inner city ([Murillo, Vayá, Romani, & Suriñach, 2013](#)). Secondly, from a data perspective it can be argued that trips, which head in and out of the city, are more accurately represented in the mobile phone location data since the average trip distance is much larger than the trips that take place within the city.

The third limitation is that trips with short travel distances (less than 10 kilometres) are excluded. This study shows that trips under 10 kilometres are prone to errors as they do not always make the user switch cell towers resulting in no log file entry being registered, which is used to detect location changes (further details on this limitation can be found in section [5.4.2](#)).

Now that the research questions and scope are defined, the next chapter continues by elaborating the research approach.

Chapter 3

Research Approach

This chapter provides all details the approach and the methods that have been used to systematically perform the actual research. Section 3.1 describes the overall approach that was used to guide this research. From section 3.2 onwards the methods used to perform the actual research are delineated. Section 3.2 gives the details on how the literature review has been performed. Subsequently, section 3.3 provides details on the assessment of the data quality of the mobile phone location data. Section 3.4 delineates the approach used to develop the trip motives model. Section 3.5 elaborates on the approach used to determine the relative importance of the factors that influence city attractiveness. Section 3.6 elaborates on the tools that have been used in this research. Finally, section 3.7 provides details on the cooperation with the external parties that were involved in this research.

3.1 Overall Approach

The approach used to systematically perform this study is [CRISP-DM](#). This is a comprehensive process model developed by [Wirth and Hipp \(2000\)](#) for carrying out Data Mining (DM) and Knowledge Discovery (KD) studies. Its popularity is underlined in a review of [DM](#) and [KD](#) process models by [Marbán, Mariscal, and Segovia \(2009\)](#) in which they said [CRISP-DM](#) to be the de facto standard for developing [DM](#) and [KD](#) studies.

This study is considered to be a [KD](#) study because the relative importance of factors that influence city attractiveness is determined using modelling techniques. To this end, [CRISP-DM](#) offers a tested, easy to understand, yet comprehensive process that ensures a structured

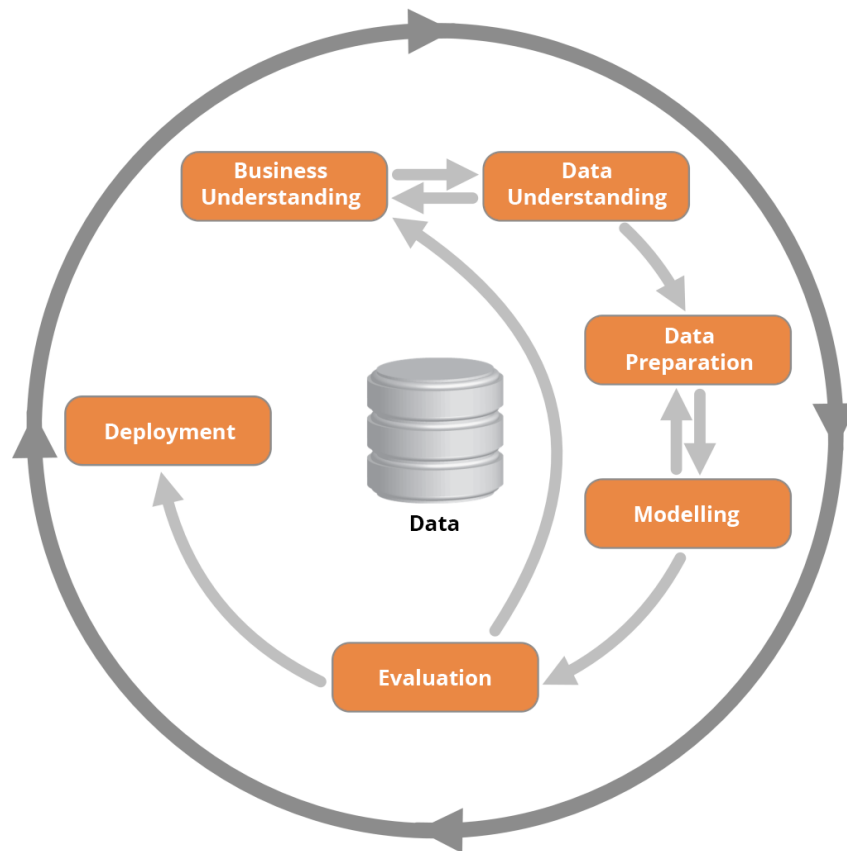


FIGURE 3.1: The **CRISP-DM** process model by [Wirth and Hipp \(2000\)](#).

approach along the way. An alternative method to **CRISP-DM** is Sample Explore Modify Model Assess (SEMMA), but since this study took place in a new domain for the researcher, **CRISP-DM** was considered to be the better choice. The reason for this is that **CRISP-DM** places more emphasis on the initial planning phases (i.e. the data and business understanding phase) than **SEMMA**. The **CRISP-DM** process consists of six phases as depicted in figure 3.1.

According to the **CRISP-DM** process model, the life cycle of a **DM** or **KD** project is broken down into six phases ([Chapman et al., 2000](#)). Within this research the first five phases of the **CRISP-DM** are performed as illustrated by figure 3.2. The last phase, i.e. the deployment phase, is omitted as this is out of the scope of this research. The subsequent paragraphs in this section give an outline of these phases. More details on the research methods that were used within each of these phases are provided by the remaining sections in this chapter.

The first phase of **CRISP-DM** is the business understanding. This phase is concerned with creating an understanding of the project's objectives and requirements from a business perspective after which they are converted into a problem definition and an research plan ([Chapman](#)

et al., 2000). Moreover, it is highly related to the data understanding phase as the data is necessary to develop a realistic plan with objectives that are feasible. The results of this phase are the content of the first three chapters of this thesis, i.e. 'Introduction', 'Research Questions' and 'Research Approach'. Within the business understanding phase of the research project a literature review was performed to identify the factors that influence city attractiveness (as an answer to SRQ1). More details on the approach used to structure literature review can be found in section 3.2 of this chapter.

The next phase of CRISP-DM is the data understanding phase (see figure 3.1). As said before the data understanding phase is closely related to the business understanding phase. Therefore, these two phases were carried out in parallel. The goal of the data understanding phase was find out what the quality is of the mobile phone location data (as an answer to SRQ2). The quality of the data was assessed (1) by performing a literature review of existing academic reports that have reported the use of mobile phone location data of Mezero and (2) by critically reviewing all steps of the process to convert the raw CDRs to the final product, i.e. mobility information. More details on the data quality assessment can be found in section 3.3. A limitation was identified in the current scaling method, which presented a potential problem for this study.

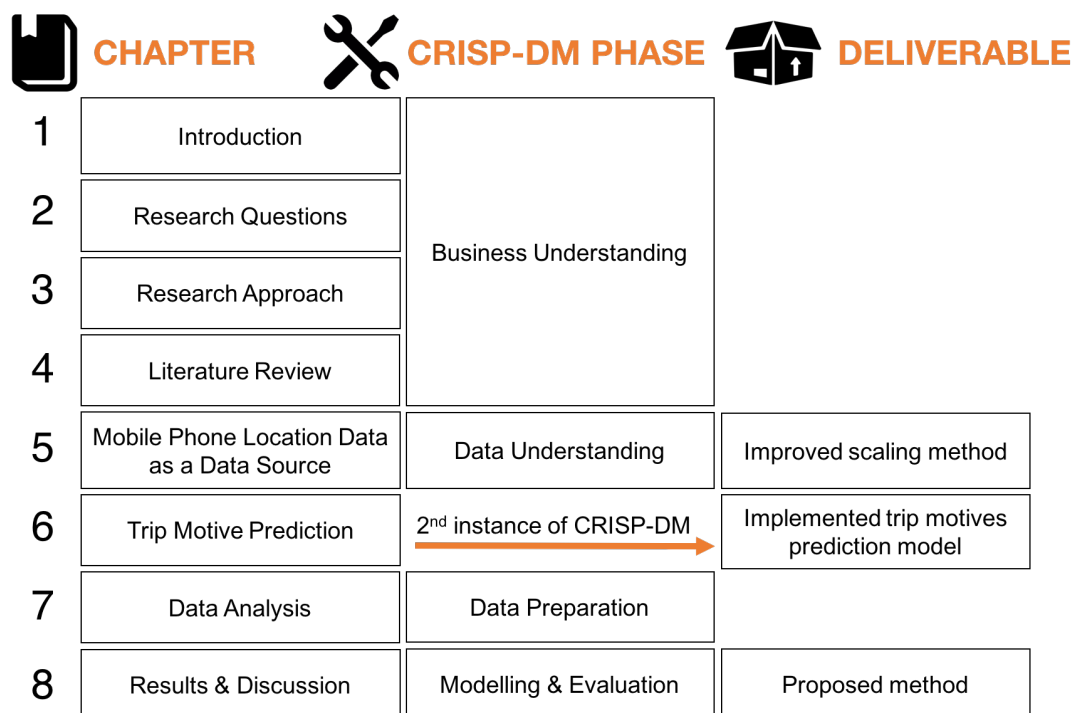


FIGURE 3.2: A schematic overview of the chapters in this thesis, the CRISP-DM phase to which each chapter is mapped and (if applicable) the final deliverable of the chapters.

Consequently, an improved scaling method was developed and applied within the context of this research resulting in more accurate data. Hence, the deliverable of chapter 5 ([Mobile Phone Location Data as a Data Source](#)) in figure 3.2 is an improved scaling method. The method was constructed and visualised using the meta-modelling technique by [van de Weerd and Brinkkemper \(2009\)](#) resulting in a Process Deliverable Diagram (PDD).

The third phase comprises all activities related to preparing the data for analysis, which can be broadly divided into (1) gathering complementary data on the factors that influence city attractiveness that were identified in the literature review in chapter 4 and (2) preparing this data for analysis.

In the modelling phase, the right method to build a model is selected and its parameters are optimised ([Chapman et al., 2000](#)). Within this research a multiple linear regression model was constructed from the prepared data. Several iterations were needed between the modelling and data preparation phase in order to test various models and to optimise the chosen multiple regression model. More details on this phase can be found in section 3.5. In the fifth phase the results of the multiple regression model were evaluated by comparing it to results from the existing literature in a discussion that is integrated in the results chapter (see figure 3.2). Finally, based on the approach used in this study a method is proposed that can be used to determine the relative importance of factors that influence city attractiveness using mobile phone location data.

3.2 Literature Review

The goal of the literature review was to answer [SRQ1](#), i.e. "What is known in literature about the factors that influence city attractiveness?". The objective of this literature review was to identify and list key factors that influence city attractiveness. There was no need to make the literature review yield exhaustive results, because the main goal of this study is not to determine *what* the relative importance of factors that influence city attractiveness is, but instead to assess how operator mobile phone location data can be used to determine this. Therefore, a narrative literature review was chosen as the preferred approach to perform the literature review.

The literature in the literature review has been collected between May and December 2015. The data source used to search for literature was Google Scholar. Snowballing, which refers to the use of the reference list of a paper or the citations to a paper to identify additional papers

(Wohlin, 2014), was frequently used as a search strategy. Due to the narrative approach of the literature review no search keywords or explicit in- and exclusion criteria have been adopted. In reporting the literature review the results of the selected papers are highlighted and its implications for this research are discussed. The literature review is concluded by answering SRQ1 and choosing which factors are included in the remainder of this study.

3.3 Data Quality Assessment

The goal of the data quality assessment was to identify the possibilities and limitations of using mobile phone location data as a data source. The approach that was used consisted of two stages: (1) perform a review of existing academic reports that have reported the use of mobile phone location data of Mezero and (2) critically review all steps of Mezero's process to convert the raw CDRs to the final product. The first stage was performed to make sure that our research is built on past efforts performed to review the data quality. The second stage was performed to obtain a better insight into how the data is formed and what the limitations are. As part of the second stage the mobile phone location data was compared to a GPS trace of a colleague (who explicitly gave permission to do so) to determine the accuracy of the mobile phone location data. The summary of this analysis is added in Appendix F.

3.4 Trip Motive Development

The attractiveness of a city differs for various target groups (Sinkienė & Kromalcas, 2010). Therefore, it was important to infer the most likely motive. The goal was to develop a model that can be used to predict the trip motive based on known trip attributes (e.g., its departure and arrival time, the day of week, whether it is to or from home, the trip travel distance). Because this a project with its own set of goals, problems and solution it has been approached as another DM project. Consequently, a second instance of the CRISP-DM process model was instantiated (as depicted in figure 3.2) to structure this project. The CRISP-DM approach is also chosen for this project for the same reasons that CRISP-DM was preferred for the overall approach, i.e. structure, comprehensiveness and emphasis on the initial planning phases. The actual chapter on trip motive prediction is structured along the six CRISP-DM phases (as depicted in figure 3.1) to reinforce the use of this structured approach.

Departure time	Destination	Transportation mode	Distance	Arrival time
6.32am	<i>to work</i>	<i>bike, train, walking</i>	28 km	7.20am
	<i>1. trainstation</i>	<i>bike</i>	2 km	6.40am
	<i>2. trainstation</i>	<i>train</i>	25 km	7.10am
	<i>3. work</i>	<i>walking</i>	1 km	7.20am
2.30pm	<i>to home</i>	<i>walking, train, bike</i>	28 km	3.18pm
	<i>1. trainstation</i>	<i>walking</i>	1 km	2.40pm
	<i>2. trainstation</i>	<i>train</i>	25 km	3.10pm
	<i>3. home</i>	<i>bike</i>	2 km	3.18pm

TABLE 3.1: A sample of the [OVIN](#) data.

To build a model on trip characteristics to predict the trip motive it is key to have a data source in which both are present. In the Netherlands the largest source of information containing both trip characteristics and trip motives is the [OVIN](#). An example of the [OVIN](#) data is presented in table 3.1. The [OVIN](#) data was prepared and a special type of decision tree called [PET](#) was chosen because it is (1) easy to translate into Structured Query Language (SQL) and (2) provides a better estimation of the distribution than the regular decision tree, and (3) requires little computation time. Subsequently, the model performance was evaluated on different levels (i.e., country, province and municipality level) using cross validation and the parameters were optimised. The final model was converted to [SQL](#) and implemented to predict trip motives using the mobile phone location data. The end result is an optimised and implemented model that predicts the most likely trip motive based on a set of trip attributes using the trips of the mobile phone location data.

3.5 Data Analysis

The goal of the data analysis phase is to identify a way in which mobile phone location data can be used to determine the relative importance of factors that influence city attractiveness. The literature review, of which the approach was presented in section 3.2, resulted in a list of factors (e.g., income levels, commercial land use, number of museums) that influence city attractiveness. The data to represent these factors was collected from trusted external sources, such as Centraal Bureau voor de Statistiek (CBS). [Wickham \(2014\)](#) proposed an approach to structure data, which was adopted due to the large amount of data that was collected for this research. It was chosen for its applicability to the R and RStudio environment, one of the primary data manipulation tools used in this research (see section 3.6).

In the modelling phase of **CRISP-DM** several statistical methods were applied, namely: ridge regression, correlation analysis, multivariate regression and multiple linear regression. In the end multiple linear regression was chosen for it is the most parsimonious, comprehensible and applicable method to determine the relative importance of factors that influence city attractiveness. However, one of the assumptions of multiple linear regression is that there is no multicollinearity between the independent variables that are included in the model. In general, multicollinearity reduces the overall R^2 and negatively influences the statistical significance tests of coefficients by inflating the variance of independent variables (Hair et al., 2006). Therefore, a **PCA** is performed which derives variables from the original dataset. In this case **PCA** is preferred over factor analysis because factor analysis assumes no multicollinearity while this is not a problem for **PCA** (Field, 2009).

Finally, to determine the relative importance of factors that influence city attractiveness multiple regression analysis is used. As long as the predictor variables are uncorrelated, the importance of each predictor is given by the R^2 -value of the predictor, and all R^2 -values of the predictors add up to the full model R^2 -value (Grömping et al., 2006).

The main deliverable of this study, a method to determine the relative importance of factors that influence city attractiveness using mobile phone location data, was constructed based on our approach used to obtain the results. The method was constructed and visualised using the meta-modelling technique by van de Weerd and Brinkkemper (2009) resulting in a **PDD**.

3.6 Tools

Over the course of this study a variety of tools have been used. The mobile phone location data was processed by the Aster platform. The output of Aster was loaded into a PostgreSQL database after which it was ready for analysis. Most of the data manipulation has been done using R, a programming language for statistical computing (R Core Team, 2014), and RStudio, a free open-source integrated development environment for R. However, this research would have taken much more time without the large number of R libraries from third party developers. The use of a library for specific parts of this research is mentioned in the respective parts of this thesis.

Microsoft Excel was mostly used for its ability to make charts. The thesis itself was written using the cloud services and $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ -processor provided by ShareLaTeX.com. The images, graphs

and diagrams (that are used throughout this thesis) were designed with Microsoft Powerpoint, Microsoft Visio, Microsoft Excel and Adobe Photoshop. The geographical images were made using QGIS and PostGIS.

3.7 Cooperation with Mezuro and Decisio

The idea for this research has originated in cooperation with Mezuro, a company situated in Weesp, which is a small city in the periphery of Amsterdam, and partner company Decisio. Mezuro¹ is a supplier of anonymous mobility data originating from mobile communication networks. For this purpose they use data from a major Dutch mobile network provider. Their partner Decisio² is a research and consultancy company situated in Amsterdam that focuses on spatial economic consultancy for governmental organisations. Next to providing data for this research, Mezuro and Decisio have provided: quality feedback, useful contacts, great solutions and challenges and a productive working environment.

This research presented a business opportunity for all parties involved, which is why they were eager to find someone with similar interests. At the start of the project it was unknown whether our goal was actually feasible. Therefore, this research presented an exploratory research opportunity for a multidisciplinary oriented academic scholar. The innovative nature of the research has made this a real data adventure and most interesting for everybody involved.

¹Additional information on Mezuro can be found at <http://www.mezuro.com>

²Additional information on Decisio can be found at <http://www.decisio.nl>

Chapter 4

Related Literature

The purpose of this chapter can be divided in two folds. First, to provide an answer to [SRQ1](#): *What is known in literature about the factors that influence city attractiveness?*. Second, to inform the reader about the factors that can influence city attractiveness.

This chapter is structured as follows. Section [4.1](#) starts with a short introduction of how the literature review is structured. The subsequent sections contain the main content of the literature review. Finally, section [4.6](#) summarises the findings and provides an answer to the [SRQ1](#).

4.1 Introduction

A literature review an essential part to determine the relative importance of factors that influence city attractiveness. Hence, without it there is no theoretical foundation to determine the factors that influence city attractiveness. Therefore, a literature review and defining city attractiveness are the first activities of this method (as depicted in figure [4.1](#)), which forms a part of the main deliverable of the present study.

The topic of city attractiveness, within the context of this research (i.e., according to our definition of city attractiveness), is multidisciplinary and highly complex. People visit a city for a wide variety of reasons (such as shopping, commuting, tourism, visiting family and relatives). Moreover, there are a lot of external factors that influence the choice for travelling (such as income, weather, accessibility of transport, and availability of alternatives). The literature on

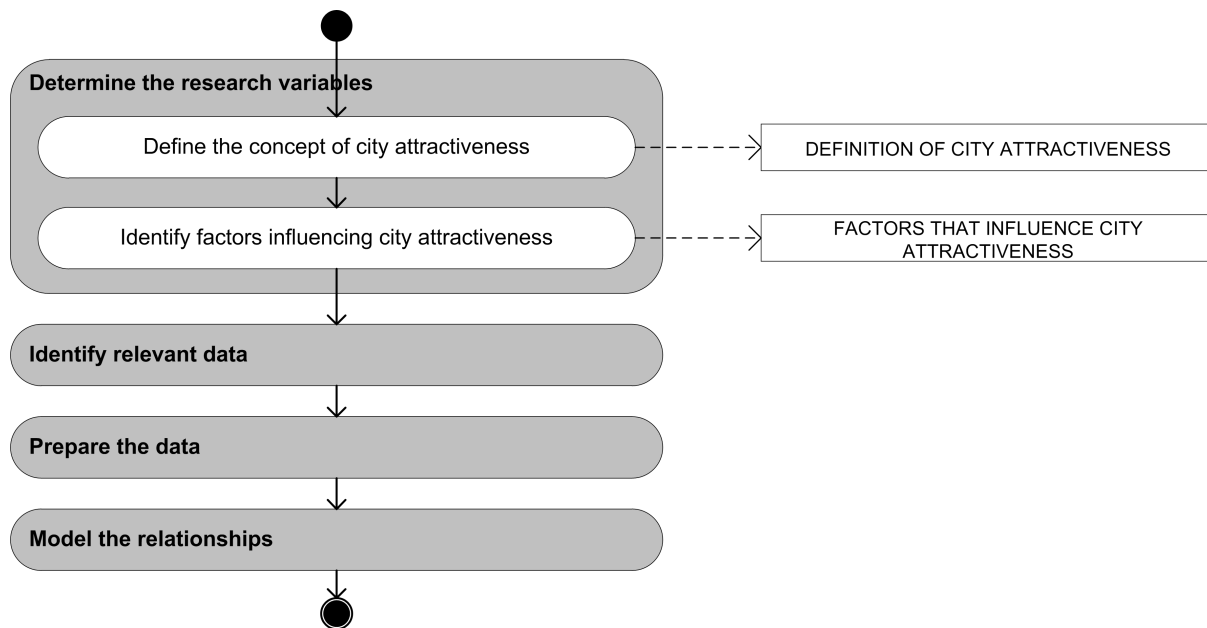


FIGURE 4.1: The first part of the proposed method is applied in this chapter. The other activities in the method are collapsed as these will be applied in chapter 7 and 8.

the factors that influence city attractiveness are spread across several different research fields. Consequently, in the next sections of this chapter, key articles are discussed from the urban competitiveness, city centre retail attractiveness, tourism and urban planning literature. This clustering is chosen based on the articles that have been found using the method described in the research approach of this study.

To keep things coherent, the findings are discussed per field of study. Each of the subsequent sections in this chapter starts with a short introduction followed by a review of the key literature within this field concerning the factors that influence city attractiveness. Finally, the relevance and implications of these key articles are discussed.

4.2 Urban Competitiveness

The great advances in information technology in the 1990s have had a large impact on the concepts of space and time (Ni, 2012, p. 14). Hence, it has decreased the average travel time from city to city and increased the speed at which we can send information between locations. This increasing trend in globalisation has not only enhanced cities' roles as the independent motors of our local economy (Sinkiené, 2008), but also increased the competition between them

(Sinkienė, 2009). Consequently, this has led to urban competitiveness being adopted as a policy goal by Europe and North America (European Commission, 2000).

Because of its adoption as policy goal the field of urban competitiveness has gained massive popularity among scientific researchers over the past 20 years. However, despite the increasing amount of attention it receives urban competitiveness remains very elusive and hard to grasp (Deas & Giordano, 2001; Kitson, Martin, & Tyler, 2004). Moreover, since it is a field still in development research about techniques to measure urban competitiveness are still lacking (Bruneckiene, Guzavicius, & Cincikaite, 2010). Therefore, it is often expressed, improved or communicated relative in comparison to other places (Malecki, 2002). For example, when expressing the competitiveness of a city like Utrecht it is often ranked as less competitive than Amsterdam but more competitive than Breda.

When looking at the concept of urban competitiveness from a more abstract point of view it can be seen as a special kind of competitiveness, a very complex and controversial concept (Delbari, Ng, Aziz, & Ho, 2015; Singhal, McGreal, & Berry, 2013). Deduced from the meaning of the parental concept of competitiveness, Malecki (2000) defined the competitiveness of places—which can be cities, regions or even nations—as the ability of local economy and society to improve the current welfare standard for its inhabitants.

Next we will discuss some key models from the urban competitiveness literature to see what we can learn from them for the purpose of this research.

Over the last decade, the increasing popularity of the field has yielded a large amount of frameworks being developed from different perspectives, indicators and methods (Malecki, 2002; Ni, 2012, p. 15). Most of them are based on Porter's (1990) diamond model of competitiveness (Healey & Dunham, 1994; Kresl, 1995). However, most of these frameworks are aimed at conceptualising urban competitiveness rather than quantifying it Deas and Giordano (2001). Thus far, a comprehensive framework is far from established because of the many aspects that are related to the urban competitiveness hindering empirical assessment (Kitson, Martin, & Tyler, 2005). Ni (2012) proposed that the theoretical frameworks on urban competitiveness can be grouped into three groups: (1) the input-outcome perspective, (2) the income perspective, and (3) the outcome perspective.

We explain how multifaceted the concept of urban competitiveness is using the city competitiveness model proposed by Sinkienė (2008). The small version of his model can be found in

figure 4.2 and shows the process of city competitiveness formation at a conceptual level. This model is based on the input-outcome perspective as can be seen by the levels within this model. The division between urban internal and urban external environment factors is important because internal city factors can be influenced by policies and urban development, but external environment factors are hard to influence.

The author recommends that in order to identify factors that influence the formation of competitiveness one must start at the external environment. Arguably, these factors are of the greatest influence on city competitiveness (Sinkienė, 2008). The external urban environment factors proposed can be grouped in political-legal, technological, economic, social-cultural and ecological-environmental. The implemented version of the model is too big to include but can be found appendix E. The internal urban environment input factors can be grouped as human factors, institutional factors, physical factors and economic factors. Each group consists of a number of indicators that say something about the group they belong to. Together they are the input for the between people, housing, work, leisure and transport that happen in a city.

The outcomes of the processes recursively influence the inputs because of the internal urban environment changes. Moreover, the level of competitiveness can be deduced from the outcomes but this value always has to be communicated relative to other cities. Sinkienė (2008) denotes

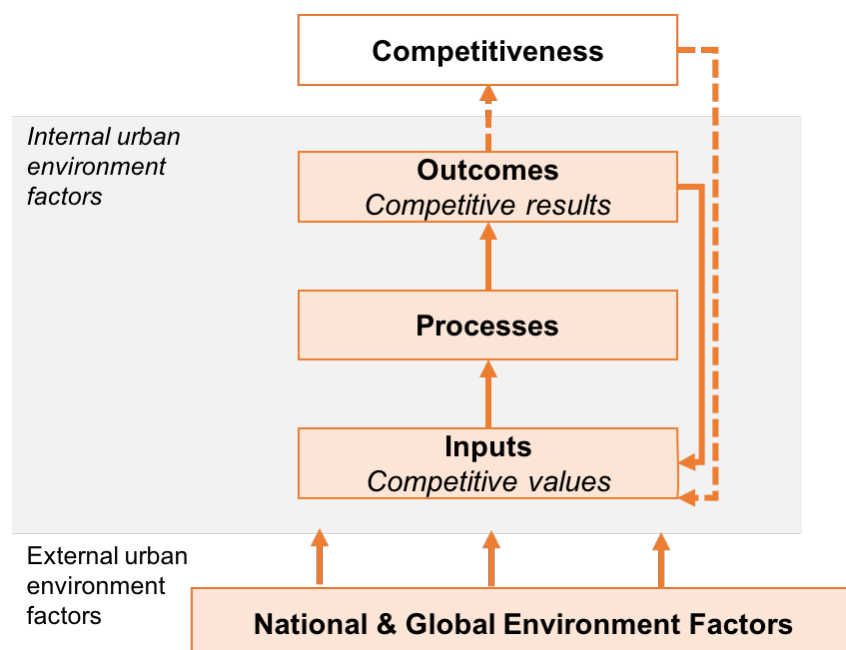


FIGURE 4.2: Conceptual urban competitiveness model (Sinkienė, 2008).

that she has not been able to find a causal relationship between inputs and outcomes. Consequently, she stresses that the individual elements in this model cannot be seen individually, because the system as a whole supplements and strengthens another making it highly complex.

The concept of city attractiveness belongs to the output of the processes within a city according to the model by [Sinkiené \(2008\)](#). What is unknown within this model, according to the authors of the model, is which inputs influence which outputs and what their relative contribution is. Therefore, we can only use their inputs as input for a brainstorm in which all indicators from the model are discussed that can logically influence the level of city attractiveness.

This model attempts to explain the determinants of urban competitiveness on an international scale, like models in this field of study. However, within the scope of this research, which is limited to Dutch cities, external urban environment factors such as political stability, IT innovation, taxes, demographics and climate are more or less the same for each city. Therefore, they are not expected to contribute to explaining the variance in city attractiveness. Consequently, external factors are, to a certain extent, not taken into account within this research and the focus will be on the internal factors of a city.

The framework by [Deas and Giordano \(2001\)](#) to assess the competitiveness of cities uses several different indicators. They assessed whether the characteristics of a city's asset base are a good representation for determining its competitive performance. The asset base they reviewed consisted of four main categories, namely economic environment, policy or institutional environment, physical environment and social environment. They concluded that the asset base within a city presents a potentially useful way of conceptualising and measuring the competitive position of cities. Interestingly, within their analysis, inner city cores were included as separate areas. They do not discuss why they did this in detail, only that the results between peripheral areas and core areas differ for the assets assessed. This makes sense, because of the high density of resources and assets in the core of a city and the availability of historic building and objects. Based on the results of [Deas and Giordano \(2001\)](#) we will use the availability of assets in the peripheral areas of a city as a factor.

Figure 4.3 is the urban competitiveness model by [Begg \(1999\)](#). It shows resemblance to the model by [Sinkiené \(2008\)](#) that we discussed previously. The sectoral trends, company characteristics, business environment and capacity for innovation in this figure are part of the internal

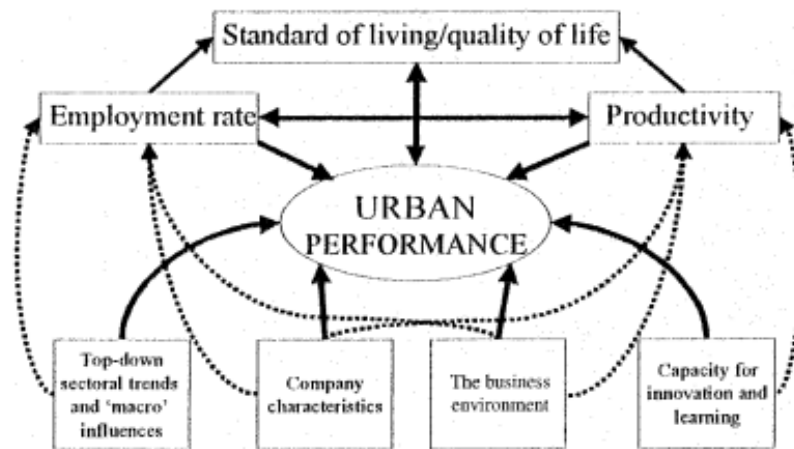


FIGURE 4.3: Urban competitiveness maze (Begg, 1999).

and external urban environment inputs in the model by Sinkienė (2008). Employment and productivity are part of the outcomes and are part of the indicators for the quality of life. Together they form the level of urban competitiveness.

In conclusion, urban competitiveness is a field of study that aims to develop a comprehensive model of the factors that influence a city's level of competitiveness. Not only the attractors within the urban environment play a role of importance but also factors that are considered to be part of the external environment such as urban policies, macro economic factors and micro economic factors, are taken into account. Consequently, we have identified that city attractiveness and the factors that influence it are part of a larger, more intricate, and highly interdependent system that can be conceptualised as urban competitiveness (referring to the model in appendix E).

However, due to the scale and complexity of this field, the amount of literature that focuses on identifying relationships between the inputs and outcomes of city competitiveness is rather limited. The takeaway from this field of study is that competitive forces influence the city attractiveness. Therefore, this factor has to be taken into account when applying the proposed method. This literary quest continues by assessing existing literature city centre retail attractiveness.

4.3 City Centre Retail Attractiveness

The increase in the number of failing shopping centres, the rising number of shopping agglomerations and increasing competition among shopping centres can be seen as some of the major trends in retail (Baker, 2006; Wrigley & Lowe, 2002). This has sparked the interest of academic researchers as well as professionals to research the determinants that influence shopping behaviour (Burns & Warren, 1995; Dennis, 2005). Moreover, with the rapid development of e-shopping there have been numerous studies which sought to quantify the performance and attractiveness of city centres' retail areas. What is surprising is that many of the popular studies regarding city centre retail attractiveness have been conducted in The Netherlands (Weltevreden, 2006; Weltevreden & van Rietbergen, 2007; Farag, Schwanen, Dijst, & Faber, 2007). This reinforces the importance of the concept of city attractiveness for the geographical application of this research, i.e. the Netherlands.

We review the research by Weltevreden and van Rietbergen (2007) who studied how the perceived city centre attractiveness in dimensions such as accessibility and convenience of shopping and range, influences the relation between e-shopping and city centre shopping. This study showed that over 20% of the online shoppers was less likely to frequent the city centre stores, due to e-shopping. However, the higher the perceived attractiveness of a city centre, the less e-shoppers are inclined to shop online and will replace their shopping with city centre shopping. Their findings affirm that as the attractiveness of a city centre increases, people are more likely to choose city centre shopping over online shopping. Therefore, by increasing the perceived city centre attractiveness people are more likely to prefer city centre shopping over e-shopping. This is interesting for our study, because we develop a method to determine the relative importance of factors that affect this level of city attractiveness. An improved understanding of the relative importance of factors that affect a city centre's attractiveness can be useful in managerial decision-making in various ways (Teller & Reutterer, 2008).

Teller and Reutterer (2008) compiled a list of factors that influence city centre retail attractiveness by reviewing existing literature. The factors are: accessibility, parking, retail tenant mix, non retail tenant mix, orientation, ambiance, atmosphere, distance, and involvement. This list of factors can be incorporated in our study as city centres are often what attracts specific groups of visitors such as tourists, day trippers and shoppers.

The last thing that was noticed when reviewing existing literature on city centre retail attractiveness is that the preferred method for assessing the perceived city centre retail attractiveness is a survey. [Khei Mie Wong, Lu, and Lan Yuan \(2001\)](#) even developed a standardised survey to assess the perceived attractiveness of shopping centres and its development over time. Although the use of a survey allows for a good assessment of the perceived importance of characteristics, it also has some limitations compared to using mobile phone location data as a method for data collection inherent to the smaller sample size. This includes the ability to make inter-city comparisons. With questionnaires you must be very aware of the fact that the results have sufficient external validity, i.e. are generalisable. By using mobile phone location data the sample size is large enough to include data on all cities and results will therefore be less prone to the limitation of external validity.

This review of the city centre retail attractiveness literature leads us to believe that there are a variety of factors that influence the attractiveness of a city centre. According to [Teller and Reutterer \(2008\)](#) these factors are: accessibility, parking, retail tenant mix, non retail tenant mix, orientation, ambiance, atmosphere, distance, and involvement. Some of these factors are hard to quantify but our list of factors should attempt to include as many of them as possible as city centres are considered to be the 'pulse of urban life' ([Borgegård & Murdie, 1993](#)) and therefore highly influential on the city attractiveness in general.

4.4 Tourism

Over the last three decades destination attractiveness has become one of the most popular research fields within the tourism research literature ([Pike, 2002](#)). Although at first glance the concept of destination attractiveness might seem no different than the concept of tourism attraction, there is a big difference between the two. According to [Krešić and Prebežac \(2011\)](#) tourism attractions can be defined as specific destination features (such as climate, landscape features, activities in destinations etc.), which have the ability to attract visitors. On the other hand, destination attractiveness refers to the cognitive image that only exist in the mind of a potential visitor of a destination. Hence, tourism attractions can be seen as the physical part of the destination attractiveness, while destination attractiveness is the cognitive image that is formed on the base of the physical attractions available at the destination ([Krešić & Prebežac, 2011](#)).

The idea behind this research has been studied before within the context of tourism research (see [Beerli & Martin, 2004](#); [Enright & Newton, 2004](#)). However, to the best of our knowledge this kind of research has not been attempted before using mobile phone location data. By evaluating whether mobile phone location data is applicable for this type of research we can further this field by providing a data source that has relatively little bias and a provides a major improvement in sample size compared to surveys.

The assumption that mobile phone location data can provide added value to the tourism research is confirmed by a study by [Ahas, Aasa, Roose, Mark, and Silm \(2008\)](#). They evaluated the applicability of mobile phone location data in the context of tourism. They found a very high correlation $R = .99, p < .05$) between the number of overnight stays in accommodations and tourist presence using to mobile phone location data, which confirms that mobile phone location data is applicable to tourism research. However, the study by [Ahas et al. \(2008\)](#) is limited to determining whether mobile phone location data is a good data source to determine the number of tourist as an alternative to tourist surveys. Hence, they did not determine what the relative importance is of factors that influence city attractiveness. Another major difference is that in this research we do not focus solely on foreign tourists, but on city visitors in general.

Within the tourism research factors that attract tourists are denoted as attractions and are perhaps the most important constituent of the tourism system ([Swarbrooke & Page, 2012](#)). They are the core of what attracts tourists to a city. This is confirmed by [Formica and Uysal \(2006\)](#) where the main principle underlying the study is that the overall destination attractiveness depends on the relationship between the availability of existing attractions and the perceived importance of such attractions.

This review of tourism literature was mostly related to destination attractiveness, which is highly applicable and related to this research. As such, two comprehensive lists of factors that influence destination attractiveness have been found. [Beerli and Martin \(2004\)](#) compiled a list of factors that affect the attractiveness of a destination. Their list of determinants that influence the destination attractiveness from the perspective of a tourist is chosen for its comprehensiveness (the list can be found in appendix [D](#)). The other list of factors was found in the study by [Enright and Newton \(2004\)](#), but this list was not used because the terminology used was too abstract.

4.5 Urban Planning

For this research, the most interesting topic within the field of urban planning is the travel demand forecasting. Travel demand forecasting is the process of estimating the transportation need, in terms of people or vehicles, in the future. The reason why we choose to include this topic in our literature review is because of the overlap it has in quantifying the attractiveness of an area using the number of trips that go in and out of an area. Almost analogous to travel demand forecasting is the "four-step" process (Martin et al., 1998). The four major steps in this process are: trip generation, trip distribution, mode choice and trip assignment (see figure 4.4), which we will elaborate on.

Trip generation is related to determining the number people or vehicle trips that are produced and attracted by each zone within the area of analysis. The number of trips produced and attracted by each zone is affected by various factors such as land use, income levels, car ownership, household size, density, type of development, availability of public transport, and quality of the transportation system (Martin et al., 1998). Apparently these factors influence city attractiveness according to our definition and should be taken account in our analysis. Using mobile phone location data we can determine exactly how many trips are generated and attracted per zone.

Trip distribution is the second step in the four-step process. In this step the trips attracted and produced are linked for each pair of zones. Although we know how many trips are produced and attracted per zone using the mobile phone location data, we do not know the purpose of these trips. This is why we review the existing trip distribution model. The gravity model is the most widely used trip distribution model (Martin et al., 1998). It is based on the laws of gravity where the force of attraction is a function of the mass of an object, the distance between objects and the gravity constant (Bruton, 1985). The problem however is that it lacks a sound

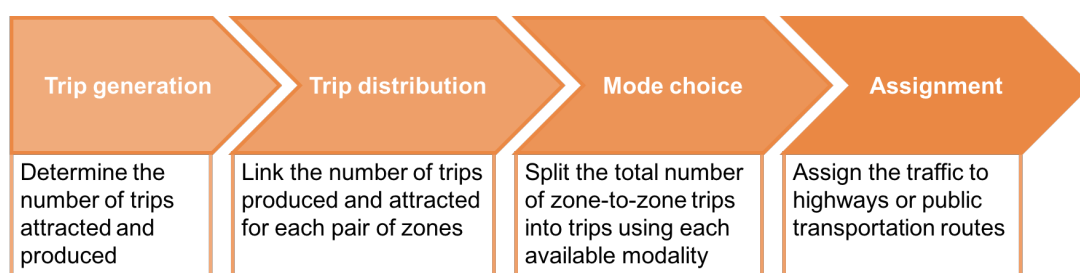


FIGURE 4.4: The four-step travel-demand forecasting process (Martin et al., 1998).

theoretical foundation (van Bergeijk & Brakman, 2010), which is exactly what we are interested in with this literature review. One way to mathematically define the gravity model is:

$$T_{ij} = P_i \left(\frac{A_j F_{ij} K_{ij}}{\sum_{k=1}^{zones} A_k F_{ik} K_{ik}} \right) \quad (4.1)$$

where

T_{ij} = the number of trips from zone i to zone j ,

P_i = the number of trip productions in zone i ,

A_j = the number of trip attractions in zone j ,

F_{ij} = the friction factor relating to the spatial separation between zone i and j , and

K_{ij} = an optional trip distribution factor to adjust for changes between zone i and j .

Most important in this formula is the friction factor, which is related to the spatial separation between two zones. The friction factor is inversely related to the distance between two zones, i.e. as the travel time increases, the friction factor decreases. From the gravity model we learn that the travel time is a factor that is inversely related to the number of trips between zones. Preferably, the travel time of a trip is taken into account as a factor.

The third step is to split for transport mode, which might yield interesting results. However, the existing method used by Mezuro to detect train and car trips is still in development and cannot be used for this research. Therefore, we will refrain from going into detail for this step. Consequently, the last step in the four-step process, i.e. assigning the traffic to highways and public transportation routes, is not applicable to this research because we have not been able to split the transportation modality.

4.6 Conclusions

This chapter set out to answer *SRQ1: What is known in literature about the factors that influence city attractiveness?*. Our literature review has shown that there are a lot of diverse factors at different hierarchical levels at play that influence the attractiveness of cities.

For this research we will use the list of determinants that influence the destination attractiveness by [Beerli and Martin \(2004\)](#) as our main list of factors for its comprehensiveness. Next to that we found that the number of trips produced and attracted by each zone is affected by various factors such as land use, income levels, car ownership, household size, density, type of development, availability of public transport, and quality of the transportation system ([Martin et al., 1998](#)). Because the list of factors by [Beerli and Martin \(2004\)](#) is from the perspective of a tourist the list of factors by [Martin et al. \(1998\)](#) is adopted as well as this list contains more generic factors that are understood to influence the number of trips going in and out of a city.

Our review of urban competitiveness literature has shown that there are competitive forces at play that influence the attractiveness of cities ([Deas & Giordano, 2001](#)). Therefore, these should also be taken into account. The factors identified by [Teller and Reutterer \(2008\)](#) that influence city centre retail attractiveness, are mostly incorporated in the list of factors by [Beerli and Martin \(2004\)](#) and are therefore omitted, except for parking fees which is included separately.

The next chapter discusses the quality of mobile phone location data, which will be used as our data source to quantify city attractiveness. It is necessary to critically review the quality of the mobile phone location data as invalid data can lead to invalid results.

Chapter 5

Mobile Phone Location Data as a Data Source

This chapter serves as an introduction and a critical review of the use of mobile phone location data as a data source. The relevance to review the quality of this data is directly in line with the [MRQ](#) of this research, i.e. determine how mobile phone location data can be used to determine the relative importance of factors that influence city attractiveness. Therefore, we must determine the quality of this data in terms of strong points and limitations, in order to be able to draw conclusions within the context of this research. In this chapter, our goal is to provide an answer to [SRQ2](#): *What is the quality of the mobile phone location data?*

The structure of this chapter is as follows. First, we describe in detail how location data can be obtained from mobile phones, which is based on how Mezero obtains location data from mobile phones, and explain how this location data can be used to infer origins and destinations of trips. Second, we assess the quality of mobile phone location data. The data quality is important because any conclusions derived within the context of this research are dependent upon the quality of this data. Third, we analyse the scaling method used by Mezero to go from the sample to the travelling population and subsequently propose an improved method to scale the sample to the travelling population.

5.1 Introduction

During the last 30 years, there has been an increasing interest by statistical institutes in novel data sources such as mobile phone location data (Daas et al., 2011). The main cause for this is their desire to alleviate data providers from the response burden with regard to providing data to statistical institutes (Snijkers, 2009). In addition, they wish to produce the census data themselves, with a sufficient level of quality and in a cost effective manner (Offermans et al., 2013). This has led to high involvement of statistical institutes such as Statistics Netherlands with research in the quality of mobile phone location data. Consequently, the majority of the existing work, mentioned in this section, is related to such institutes.

First and foremost, however, is the question of whether mobile phone location data can provide anything useful. In recent years, studies in various countries have showed that mobile phone location data can be used as a data source to get a better insight in behaviour, social networks and mobility patterns of the masses (Ahas et al., 2008; Eagle, Pentland, & Lazer, 2009; Becker et al., 2011; Palchykov, Kaski, Kertész, Barabási, & Dunbar, 2012).

The results of the study by Offermans et al. (2013) in particular are very interesting, because they specifically mention the use of the Mezero dataset for the purpose of their study. However, the mobile phone location dataset used in their study dates back to early 2013. Hence, the results they published can be irrelevant in view of recent development and changes in the algorithms of Mezero over the last 3 years. Therefore, we will need to verify that their results are still valid. However, before discussing more existing literature we will describe how location data can be obtained from mobile phones.

5.2 Obtaining Location Data from Mobile Phones

There have been a number of students who have worked with the data from Mezero (Keij, 2014; Van Kats, 2014; Brouwer, 2015). All of them have used the mobile phone location data from Mezero for their research. Consequently, all of them have had to describe the method used by Mezero to elicit location data from mobile phones. These past efforts combined provide an extensive description of how location data can be obtained from mobile phones.

Reiterating these elaborate descriptions in detail would not yield any scientific value. Therefore, this section provides a relatively brief description of how Mezero obtains location data from



FIGURE 5.1: From left to right: a cell tower with antennas for various frequencies (2G/3G/4G), a camouflaged cell tower, and a cell tower located on top of a building.

mobile phones. Our goal in this section is limited to providing the reader with the necessary understanding to interpret this research. For the most detailed description and an overview of existing techniques on how to obtain location data from mobile phones we refer to the research by [Keij \(2014\)](#).

5.2.1 The Physical Telecom Network Infrastructure

The telecom network has been set up to enable mobile devices to communicate. To enable mobile devices to communicate, cells are attached to a cell tower or a high building (see figure 5.1), called the cell site, and provide signal to devices within range. The area that a cell provides service to takes the shape of a two dimensional cone. The service area is controlled by the following parameters: angle, radius, direction and location (see figure 5.2). The location, direction, angle and radius of all cells are registered in the cell plan. Table 5.1 shows the cell plan table. Each row corresponds to a single cell. The first column contains a unique identifier for each cell (the cell code) and the columns thereafter contain the parameters of each cell. These cell parameters can be adjusted by the network operator according to fluctuations in the demand of network capacity. Cells can only provide coverage to a limited amount of devices at the same time. A practical example of this limitation can be experienced during New Year's

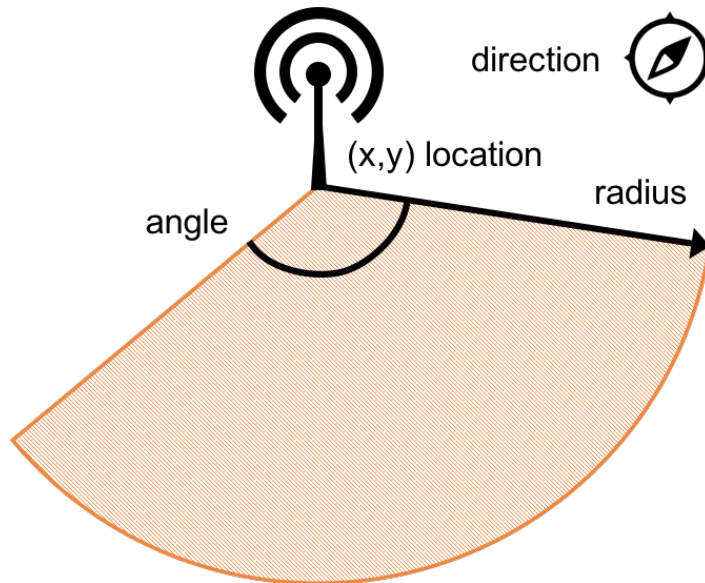


FIGURE 5.2: The adjustable parameters of a single cell located on a cell tower.

Eve. At that time it is hard to make a phone call due to people who are using their mobile phone simultaneously. Cells can be repositioned, removed or re-adjusted according to the needs. Therefore, Mezuro receives and processes a revised version of the cell plan every day.

Next to the cell plan there are events, which are generated by Mobile Stations (MSs) using the network. A *MS* is a device that contains a Subscriber Identification Module (SIM). The majority of the *MSs* are mobile phones (Van Kats, 2014). However, there are also tablets and dongles that contain a *SIM*. The location of a *MS* varies over time. Conversely, the location of the cells remains fixed during the course of a day as specified by the cell plan. Requests from the *MS* such as Short Message Service (SMS), voice calls and data induce network activity, which in turn leads to events being registered by the network operator. In other words, every time a user makes a call, sends a *SMS* or uses data, the network provider checks the balance of the user and saves the details of this request as an event in the *CDRs*.

cell_code	radius	direction	angle	tower_x	tower_y
0012027349	7070	90	63	120470.686039663	495345.950047
0012027358	7070	210	63	120470.686039663	495345.950047
0012027359	7070	330	63	120470.686039663	495345.950047
0012031044	4949	60	65	114399.609406567	518147.79938571
0012031045	4949	180	65	114399.609406567	518147.79938571

TABLE 5.1: A part of the cell plan table, which contains the parameters of every cell.

event_datetime	hashed_id	event_type	direction	country	cell_code
2015-01-01 07:28	66481215	V (Voice)	OE (Originating End)	204	0007332832
2015-01-03 13:49	48404065	V (Voice)	TE (Terminating End)	204	0004947382
2015-01-03 05:35	48404065	V (Voice)	OE (Originating End)	204	0000150641
2015-01-04 17:48	60909585	D (Data)	U (Direction Unknown)	204	0013635671
2015-01-01 03:35	09064844	V (Voice)	OE (Originating End)	204	0012210797

TABLE 5.2: An example of the CDRs that are processed to obtain mobile phone location data.

Table 5.2 shows an example of the CDRs and contains the following: the time when the event was registered (`event_datetime`), the encrypted phone number (`hashed_id`), the type of event (`event_type`), whether the event was a closing or opening (`direction`), the country code (`country`) and the unique identifier of the cell that handled the request (`cell_code`). The location of the MS at a specific point in time can be estimated by matching the cell code from the event to the cell code in the cell plan. In the remainder of this report we will interchangeably use the terms 'user' and MS to refer to a unique device that is present within the sample in order to improve the readability of this report.

5.2.2 From Locations to Origins and Destinations

An example of the cells that handle network activity during an average trip is depicted in figure 5.3. The route that was travelled is plotted using GPS coordinates that were obtained from an application that collects GPS data from the mobile phone of the user. As can be seen, the cells vary in location, direction, radius and angle. The technique used by Mezuro heavily relies on the size of the area that a cell provides service to. Hence, the bigger the service area of the cell, the less accurate the location estimation.

What can also be seen in figure 5.3 is that the cells tend to overlap each other. This feature of the data can be used to improve the accuracy of the location estimation. To go from the CDRs to location estimations Mezuro first calculates so-called "frames" from the CDRs. A frame is the service area of a cell in which the MS has been for a certain period of time. When there is overlap between service areas of the cells (as seen in figure 5.3), the frame generation algorithm will use the cell with the largest service area and ignores the smaller cells until the MS has an event with a cell that does not overlap the previous cell. At that time a new frame is generated and the process continues all events are processed.

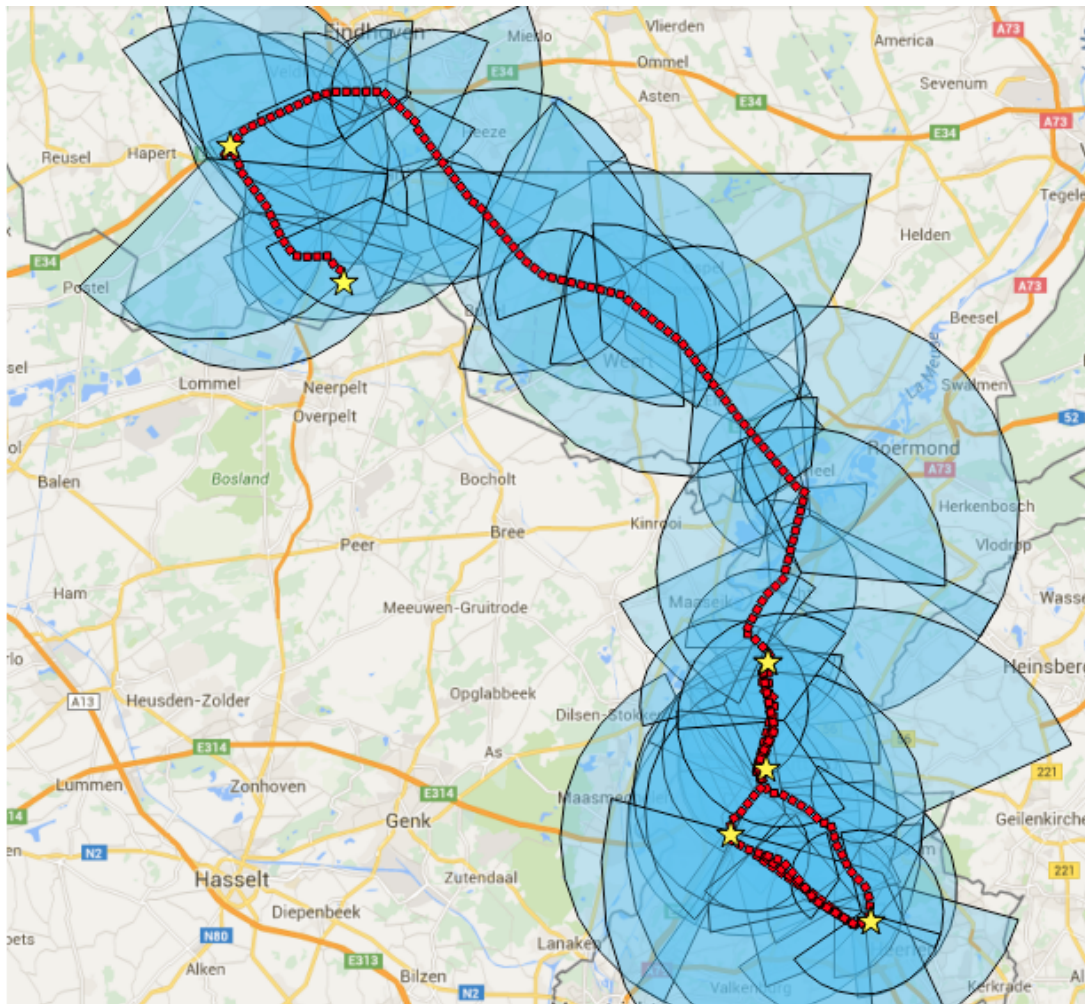


FIGURE 5.3: The cells that are connected to when travelling the red dotted line. A destination is defined as a location where a user stays for 30 minutes or more and is depicted as a star. The cells are obtained from the mobile phone data, while the route in red is obtained from GPS data.

The frame generation algorithm reduces the number of data points. The frames are used to produce an OD matrix, which among others contains information about the number of trips between places. Mezero defined a destination as a place where a MS is present for 30 minutes or more. However, since this was defined arbitrarily we compared GPS data to the data from the OD matrix and noticed that there were often inconsistencies between the two in origins, destinations and travel time. For example, more than once we saw a person travel from Utrecht to Weesp by train, which should take about 35 minutes according to the GPS data and our personal experience. However, the OD matrix from the mobile phone location data showed that the person made a trip from Utrecht to Hilversum, which is a place in between Utrecht and Weesp, in one minute, stayed in Hilversum for 28 minutes and then travelled from Hilversum to Weesp in 1 minute. When comparing the GPS data with the OD matrix data, it can be

seen that the intermediate destination and travel times are incorrect. If we would proceed our research using this data, these results would have a negative impact on any further analyses we would perform using this data.

This gave us reason to solve this problem. We found that the large cells were troublesome as it happened frequently that people were in the service area of a single cell for 30 minutes or more, making it a destination instead of a trip. Moreover, since most of the time was being assigned as the duration of stay at the intermediate destination, the time that was left for travelling from and to the invalid intermediate location was equal to a couple of minutes. Therefore, the travel time had many incorrect values. We decided to compare several solutions to solve this problem. One of them was to remove the cells with a large radius. However, removing cells with a large radius would imply that a certain percentage of the events would be invalidated, as they are handled by these cells. This meant that there is a trade-off between the number of events and the accuracy of the location estimation. To get a feeling for this trade-off we analyse how many events are still available when deciding to remove cells with a radius greater than a certain value. This is shown in table 5.3. From the data in this table we conclude that at approximately 12.5 kilometres the decrease in the number of events is dismissible. Hence, we choose 12.5 kilometres as a cutoff point leaving us with 94% of the events to estimate the origins and destinations from. The details and results of this analysis can be found in Appendix F.

About 6% of all events are ignored because they are handled by cells with a radius greater than 12.5 kilometres. We noticed a large concentration of trips with travel times under 5 minutes. Therefore, our goal was to reduce the number of trips with incorrect travel times. The exclusion of cells larger than 12.5 kilometres led to an improvement in the travel times as the number of trips with a travel time under five minutes decreased from 9.2% to 6.9%. Moreover, the coverage in the Netherlands is such that by excluding cells larger than 12.5 kilometres the coverage remains unaffected. We can add that the current OD matrix algorithm favours the use

Cell radius up to (km)	Percentage of events
2.5	60%
5	77%
7.5	86%
10	91%
12.5	94%
15	95%

TABLE 5.3: The number of events handled by cells up to a specific radius.

of smaller cells as smaller cells have less uncertainty in determining the location of the MS and result in more accurate location estimations.

5.3 General Data Characteristics

By analysing the general data characteristics of the mobile phone location data from Mezero, we hope to provide an overview of the size and scope of the dataset and highlight trends that exist within the data. At the time of writing about 370 million events are generated per day by 3 million subscribers. However, the first thing that has come to our attention when analysing these numbers, is that the number of events produced on a daily basis has been increasing steadily over time (see figure 5.4). This is due to the increasing number of subscribers who are switching to 4G technology, which produces about 5 times more events than 3G technology. This trend is beneficial for the data quality, because the core of the location estimation algorithm is based on these events. Therefore, more events have a positive effect on the accuracy of the location estimations.

The reason is that more events make it more likely that an event will be created near the time of departure or arrival of a person. Hence, more events should give a more accurate representation

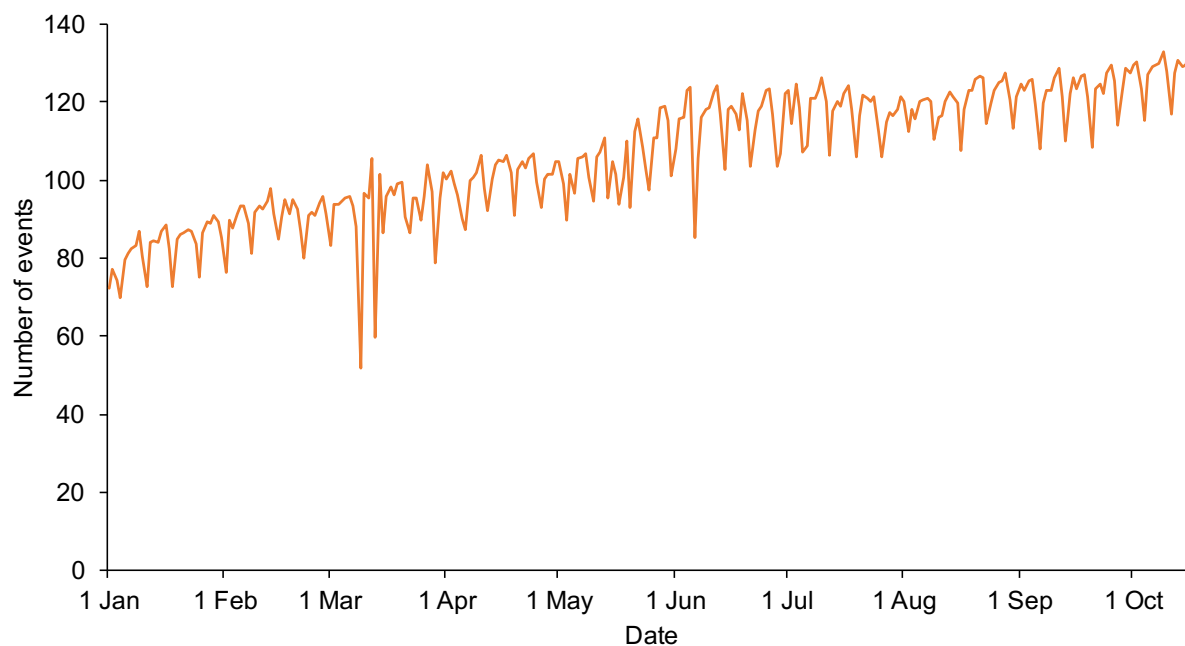


FIGURE 5.4: The average number of events per user in 2015.

of trip attributes such as the actual time of departure, actual time of arrival, and the actual length of stay at a destination.

In addition, there are some drops in the average that require explanation. First of we must note that the exact details of when events are created are not known in detail, not even at the service provider itself as the network is mostly outsourced. However, from our analysis in appendix F we are aware of a positive correlation between phone usage and the number of events created. Thus, generally speaking, the more a phone is used the more events are created.

Looking at figure 5.4 again, there are small drops that occur about four times per month. These drops are all Sundays where decreased user activity leads to less events being generated. The three bigger drops in the average number of events happening mid March and early June are related to incorrect data processing by Mezero. Even though this does influence the quality of the data negatively as it becomes less reliable, it has no implications for the remainder of this research. The reason being that the data used for the analysis only concerns October 2015, which after performing a similar analysis, is free of erroneously processed data. The last thing that we like to highlight is the almost complete absence of the effects of national holidays. The reason that they are barely visible in this graph is that the graph is a ratio between the total number of users and the total number of events. When we look at these statistics independently, the decline during holidays is visible in both. However, when both decline at the same rate it means the ratio observed in figure 5.4 remains unchanged.

To conclude, the negative effects observed by looking at descriptive statistics of the data are irrelevant as they occur outside the period that is analysed within the context of this research. On Sundays, however, less events are generated, but this error is systematic and directly proportional to the number of trips being performed. Therefore, the quality of the results of the trip generation algorithm remains unaffected during the course of a week. The positive effects observed are relevant as the increased adoption of 4G technology by the majority of the subscribers leads to more events being produced, which in the long run leads to more accurate location estimation.

5.4 Limitations

According to the authors the limitations that affect the mobile phone location data can be divided into two main categories. The first category consists of limitations that are due to

privacy regulations, while the second category of limitations are due to the technical specifics underlying this data source. The subsequent sections extensively discuss these limitations and elaborate upon the consequences these limitations have for this research project.

5.4.1 Limitations due to Privacy Regulations

Privacy and data protection is a key aspect of mobile phone location data as formalised by the Dutch Data Protection Directive (in Dutch this law is called "Wet Bescherming Persoonsgegevens") (Ahas et al., 2008). Before the mobile phone location data can be analysed, three measures need to be taken in order to satisfy these privacy regulations.

Firstly, data at the level of the individual can only be processed at the servers that belong to the service provider. The unique identifier, i.e. the phone number, that can be used to track a device is encrypted when this data is sent from the service provider to Mezero. Consequently, tracing individual persons will be much harder.

Secondly, the phone numbers are encrypted using one-way hashing. One-way hashing makes use of a one-way function that is easy to compute but whose inverse function is computationally intensive to compute (Lampport, 1979). The choice for one-way is made simply because it should be infeasible to convert the unique identifier to a phone number again. The hash function uses a hash key to generate the pseudo-random result. Pseudo-random refers to the fact that the same result can be generated again using the same hash function, hash key and input. By using a new hash key the unique identifiers can have a lifespan that is in accordance with the Dutch Data Protection Directive, which is equal to one month. This implies that the unique identifier, used to identify a user, changes every month. Hence, a user can be traced for the maximum duration of one month. After a month each user is assigned a new unique identifier due to the changed hash key.

Thirdly, the output is always an aggregate of multiple users. To prevent any unauthorised person from viewing the data of an individual user, at least 15 users should be involved in each aggregation. Aggregations which involve less than 15 persons are omitted. That is because mobility patterns from mobile phone data are so unique that only four spatio-temporal records are necessary to be able to identify 95% of the users according to de Montjoye, Hidalgo, Verleysen, and Blondel (2013). This has implications on all instances where in the output less than 15 users are aggregated. Most notably, this has implication in for example rural areas, with very

little activity, and in all areas where short time periods are selected, as these data are prone to being omitted. When these three measures are combined they ensure that it is difficult to identify an individual person using this data.

Logically, the three measures that are taken to satisfy privacy regulations cause limitations for the analysis of the data. Depending on the type of analysis being performed any of the three measures can have the most impact. Therefore, during the course of this research it is essential to keep these measures in the back of the head because they can explain outliers or unexpected results in analyses.

Using figure 5.5 we elaborate on the process of how the mobile phone location data is created and how anonymity is guaranteed in analyses performed by Mezuro or any of its clients. When looking from left to right the first object is the origin of the data: the **MS** of the user. As we explained earlier in this chapter, whenever a mobile phone requires a signal, for either calling, texting or data, it attempts to connect with a cell within range that is located on a cell tower. When this happens the identity of the subscriber is validated and when it checks out a connection is established. At the same time the service provider saves this event as a **CDR** in a database.

This database is used by the provider for billing purposes. The **CDRs** contain, among others, the unique identification code that corresponds to the cell that the mobile phone connected with. Using the cell plan this cell's identification code can be related to the the location of the cell on a map and its other properties. These two variables combined give an approximate location of the user at that point in time.

In order to process these large amounts of data conventional systems do not suffice. Therefore, a system capable of processing large amounts of data (often referred to as big data) is depicted as a black box. The reason it is depicted as a black box is to emphasise that it is impossible

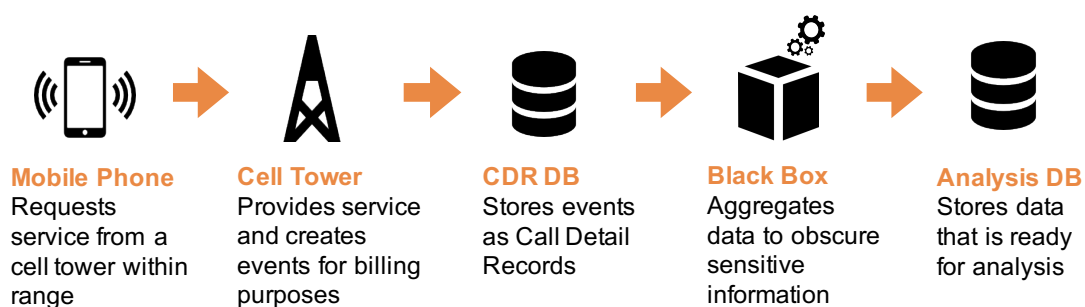


FIGURE 5.5: The data anonymisation process.

to see the data that it processes exactly; only the queries that provide the instructions and the outputs produced by the system are visible. This ensures the privacy of the users whose mobility patterns are represented by the data (de Montjoye et al., 2013). The black box is also the place where the phone numbers are hashed and where the data is aggregated. Once the query has finished running on the black box the aggregated and anonymised data is transferred through a secure connection to Mezero and saved in a local database, ready to be analysed.

5.4.2 Technical Limitations

Next to limitations that are due to privacy regulations there are also limitations that exist due to the technical specifications inherent to this data source. For instance, there is reason to believe that small travel distances (distances under 10 kilometres) do not always go out of range of a cell, thereby failing to establish a connection to a new cell within range. Because it does not always happen that a user connects to a new cell, and his movement is registered, the mobile phone location data at lower travel distances can be biased. For example, users in cities are more likely to have 5 kilometres trips registered due to smaller cell radii in cities than similar trips in rural areas. With longer travel distances (distances greater than 10 kilometres) this bias is much less likely to occur as these distances cause a device to go out of range of the first cell, causing new connection and events to register at a new cell. To determine the size of this problem the number of trips per distance class from the mobile phone location data can be compared with a representative sample of the travelling population.

Throughout this chapter we explicitly refer to the travelling population to emphasise the fact that the sample is a representation of the part of the population that travels and not of the total population (i.e., the total number of inhabitants). Moreover, the travelling population is a subset of the total population. Hence, the travelling population will always be smaller than the total population because there will always be people who do not travel. Examples for not travelling include, the immobility of senior citizens, the unwillingness, or the dependence on other persons to travel.

The representative sample of the travelling population to which we compare the mobile phone location data is obtained from OViN. OViN is a large scale enquiry where people are asked to extensively log the details of the trips they make during a single day (CBS, 2014). In this chapter the data quality of OViN will not be discussed in depth, because this chapter is limited

to a discussion about mobile phone location data. For a quality assessment of OViN we refer to chapter 6, in which OViN is used to build a model to predict the most likely trip motive.

Figure 5.6 depicts a graph that represents the number of trips per travel distance according to two different independent data sources, namely OViN and mobile phone location data. Within this graph the number of trips of both data sources has been indexed on 26 kilometres, i.e., set to 100% at a travel distance of 26 kilometres. In the OD matrix Mezuro omits trips that originate and end within the same Mezuro area in order to be consistent with their lowest level of granularity. Therefore, we argue that the maximum diameter of all Mezuro areas, which is equal to 26 kilometres, should be used as the baseline for which all trips are detected. Moreover, we know that the maximum cell radius is 12.5 kilometres, because we omitted the cells greater than 12.5 kilometres as previously explained in subsection 5.2.2. On top of that, we know that cells can have an angle of 180 degrees. So if we want to know the maximum trip distance a user can make without changing cells we double the maximum cell radius, which corresponds to the maximum travel distance of 25 kilometres. We take the largest of the two, which is 26 kilometres, and set this as the baseline where all trips are detected.

Figure 5.6 shows that the trips with small travel distances are not reliably detected (i.e., less than 10 kilometres). This is because the broadcasting range of cells can range from anywhere between 10s of meters up to 12.5 kilometres. The radius, angle, direction and location of a cell are known, which can be used to estimate where a user is present at a specific point in

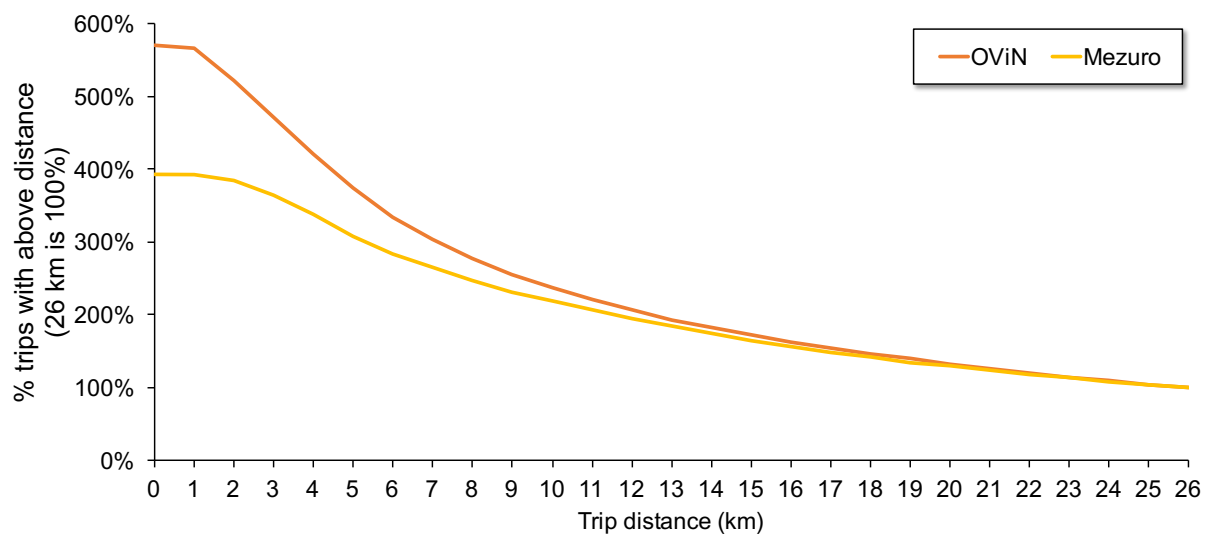


FIGURE 5.6: Comparison between OViN and the mobile phone location data of the relative number of trips per travel distance. Both datasets have their number of trips set to 100% at a travel distance of 26 kilometres.

time. Since any movements within this area can, but do not necessarily have to, induce the user to switch cells, it is uncertain if the movement is logged in the events saved by the operator making the error random. One thing that can also be noted from the graph is that the shorter the travel distances are, the less likely it is to be represented in the mobile phone location data. Hence, shorter travel distances are less likely to make a user change cells. From travel distances of 10 kilometres and upwards the number of trips is, according to us, within the acceptable limits of [OVIN](#). Therefore, this research will only use trips with travel distances greater than 10 kilometres.

5.5 Scaling Devices to the Travelling Population

The mobile phone location data should be a sample that is representative for the Dutch travelling population. We review whether that is the case by (1) analysing the current method that Mezero uses to go from devices to the travelling population and (2) by assessing how well the data represents the Dutch travelling population. The representativeness is assessed by looking at the demographic and geographic dimensions of the data. Finally, we propose a method to improve some of the shortcomings of the current method.

5.5.1 The Current Scaling Method

In order to go from users to the travelling population the data should adequately represent the travelling population. In other words, to go from the sample to the travelling population a scaling factor should be calculated and applied.

In order to do this, Mezero currently performs two steps ([Geerts, 2014](#)). First, a baseline population is determined to compensate for people being abroad, on a holiday or on a business trip. The procedure they use is to select the week where the maximum number of unique devices are detected in March or November. During these two months the least number of people are away from home, according to tourism statistics, making it the best option available to set a baseline ([Geerts, 2014](#)).

Second, a correction is applied for regional deviations in the use of mobile phones. To do this Mezero calculates a scaling factor based only on the geographic dimension of the data. This means a scaling factor is calculated by determining the penetration of mobile phone subscribers

for each Mezero area. In practice they calculate the number of active users per Mezero area and divide that by the number of inhabitants per Mezero area. An active user is defined as a user that has an event for that specific day somewhere in the Netherlands. Active users are assigned to the Mezero area in which their home address, which is inferred by Mezero based on their most frequent overnight location during a month, is located. Arguably, this ratio provides a good indication to correct for some inconsistencies in the use and possession of mobile phones.

With a view on using this method to go from the sample to the travelling population within the context of this research, we will continue our analysis by looking at the geographic and demographic dimensions of the data in the following two subsections.

5.5.2 Geographic Representativeness

Here we analyse what the size is of the inconsistencies within the geographical dimension of the mobile phone location data. [Offermans et al. \(2013\)](#) researched the mobile phone location data of Mezero and concluded that the market share of the mobile phone provider varies a lot per area. However, they concluded this based on comparing the demographic details of the subscribers to the population. We do not have access to the demographic details of the subscribers and therefore we perform our own analysis. To determine how well the mobile phone location data is represented spatially, the first thing we do is analyse the scaling factor as it is currently being used. The scaling factor is applied to correct for the variation in geographical spread of the number of subscribers per Mezero area and is calculated per day. The scaling factor is the ratio of the number of inhabitants compared to the active users in that area, which basically represents the penetration of subscribers per Mezero area. Mezero areas are used as they form a good trade-off between the accuracy and reliability of location estimation. Moreover, they are for the most part compatible with postal codes and municipality borders and can therefore easily be converted.

Figure 5.7 depicts the average scaling factor per Mezero area. The scaling factor is the result of the two steps explained in the previous subsection 5.5.1. What can be noticed is that the scaling factor varies a lot nationally. The categories are chosen in such a way that they each contain 1/6th of the total number of Mezero areas. Therefore, we conclude that two third of the Mezero areas have an average scaling factor between 4.29 and 7.43, during the first nine months of 2015. This means that the penetration of subscribers within two third of the Mezero areas lies between 13.3% ($100\% / 7.43$) and 23.8% ($100\% / 4.29$) of the travelling population.

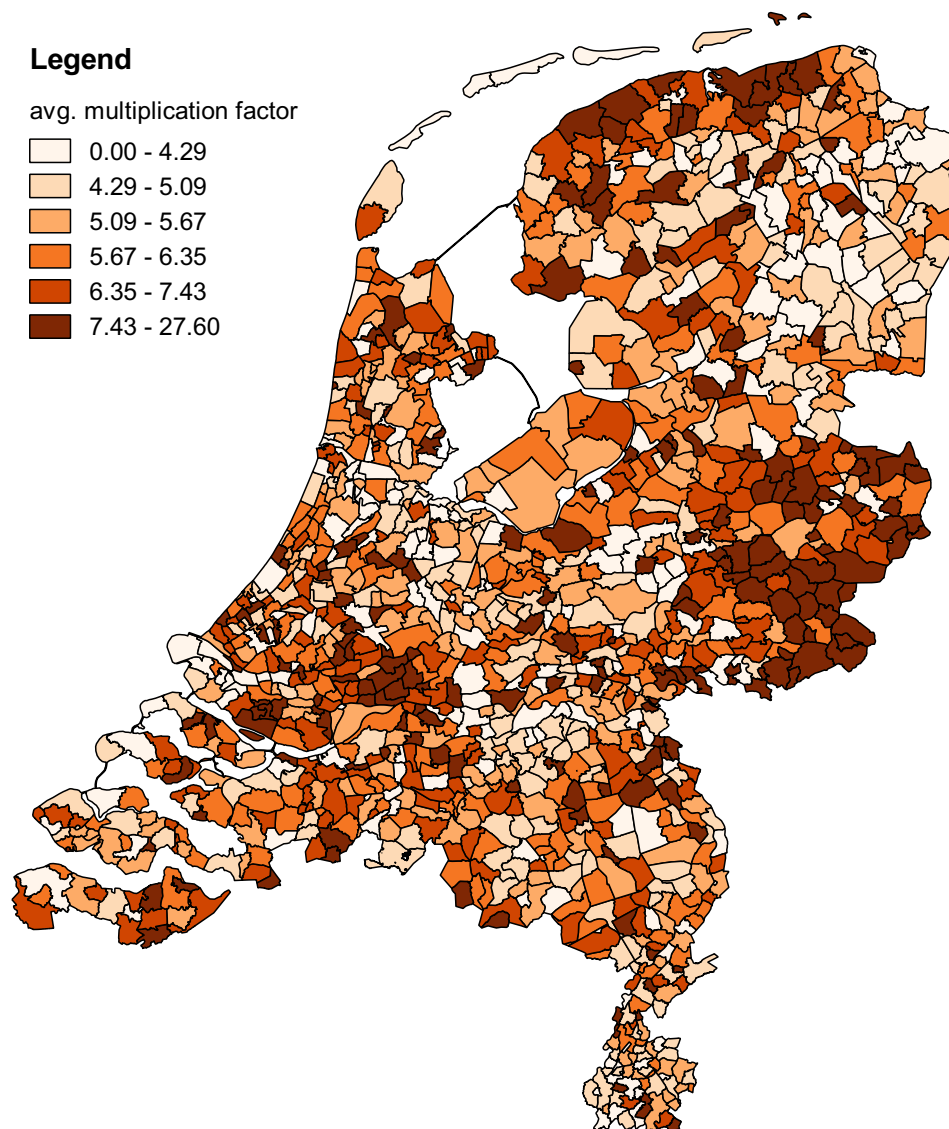


FIGURE 5.7: Average scaling factor per Meuzero area over the first nine months of 2015.

Note that this differs from the market share of the provider, because we divide by the number of inhabitants and not the number of mobile phones.

Because the scaling factor is calculated per day it is possible to see how it changes over time. We quantify this change by calculating the standard deviation of the scaling factor. However, when we look at the standard deviation on its own it does not tell how many people are affected. To take that into account we correct for the number of active users in that area by dividing the standard deviation by the average scaling factor. These results are depicted in figure 5.8. The categories here are also chosen in such a way that they contain an equal number of Meuzero areas.

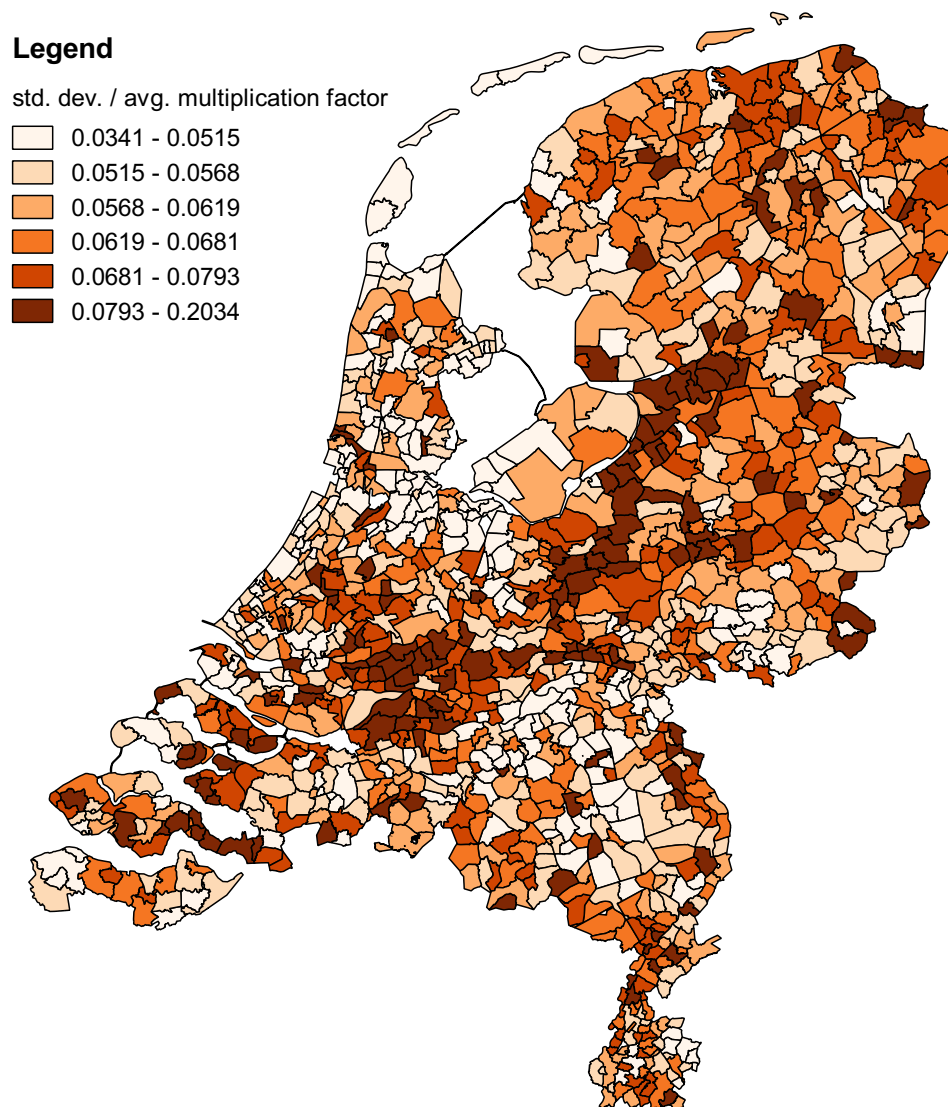


FIGURE 5.8: Geographical variation of the scaling factor corrected for the travelling population.

Most interestingly, the highest standard deviations form a characteristic pattern that highly resembles the "bible belt". The bible belt is a nickname for a collection of highly religious areas within The Netherlands. Further research confirms that when isolating this effect, by just looking at the standard deviations of the scaling factor on Sundays, the effects are even more prevalent. This behaviour is most likely explained by religious people turning off their mobile phones on Sundays for religious purposes.

5.5.3 Demographic Representativeness

The mobile phone location dataset of Mezuro should be an unbiased representation of the mobility patterns of the travelling population of the Netherlands. The current method used by Mezuro does not correct for any of the demographic differences between the travelling population and the subscribers in the sample.

[Offermans et al. \(2013\)](#) did a study to evaluate the demographic representativeness of the users in the mobile phone location data in this study. [Offermans et al. \(2013\)](#) compared the demographic details of subscribers such as age and gender, with demographic details from the municipal population register (in Dutch: Gemeentelijk Basis Administratie). Figure 5.9 shows the penetration of the sample in the Dutch population. However, they only have the demographic details of users that do not have a prepaid or business contract, which means the real distribution might differ. They conclude that young people might be underrepresented in the data, because these young people might use phones that are contracted to one of their parents' name. The overall demographic representativeness over the Dutch population appears to be good ([Offermans et al., 2013](#)).

However, we are less interested in how the data relates to the population in general and are more interested in how the data relates to the travelling population. For example, children might be less likely to travel and consequently are less likely to show up in the mobile phone location data. We need a scaling factor that takes the chance that a person of a certain age group makes a trip. Hence, we propose to use a new scaling factor that incorporates this. In the next section we elaborate on this scaling factor.

5.6 Improving the Scaling Method

In the previous sections we discussed the current method that Mezuro uses to scale the sample to the travelling population. We found that the overall demographic representativeness of the data is good. However, the current method to scale the sample to the travelling population shows several limitations. First, the penetration of mobile phones differs per age group, which is not taken into account. This will affect the likelihood that parts of the travelling population are seen in the mobile phone location data as some of the age groups simply don't carry around a mobile phone during their daily activities. For example, a child of 5 years-old is much less

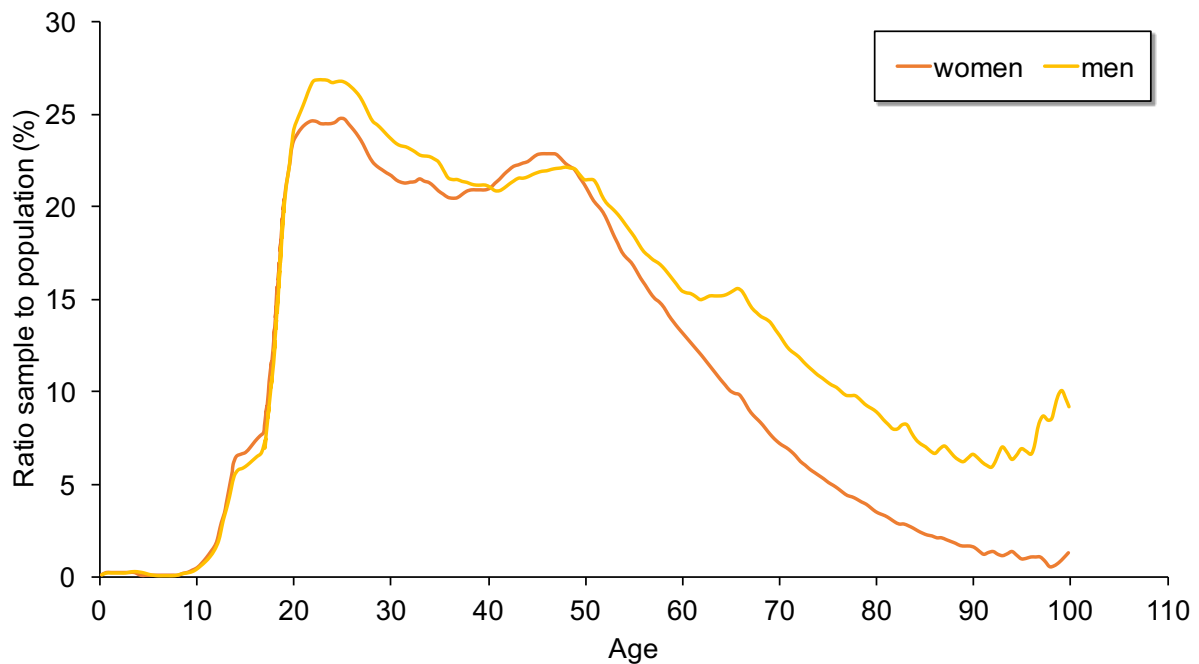


FIGURE 5.9: The penetration of subscribers per age group per gender (Offermans et al., 2013).

likely (4%) to possess a mobile phone than an adult aged 25 (96%) (Ofcom, 2014; Telecompaper, 2015). Therefore, we need to determine the penetration of mobile phones per age group and apply a correction to the data.

Second, in [OVin](#) we found that the likelihood that a person makes a trip greater than 10 kilometres differs per age group, which is not taken into account in the current scaling factor. This will also affect the likelihood that parts of the travelling population are seen in the movement data as some of the age groups simply do not perform these trips as frequently as others. This likelihood of a person travelling over 10 kilometres is depicted in figure 5.10. For example, this figure shows that a child aged 5 is less likely to make a trip greater than 10 kilometres than a person aged 40. Hence, this implies that children aged 5 are less likely to show up in our dataset than people aged 40. Moreover, it can be seen that young children are more than twice as likely to make a trip of over 10 kilometres during the weekend compared to workdays. A possible explanation would be that these children accompany their parents on family visits during the weekends. To take this variation into account we need to apply another correction in order to adjust for the differences in the likelihood that a person travels.

We will now elaborate on the choice of dimensions, i.e. age, Saturday, Sunday, workday and workday holiday. By aggregating 5 years per age group most of the variation is included while

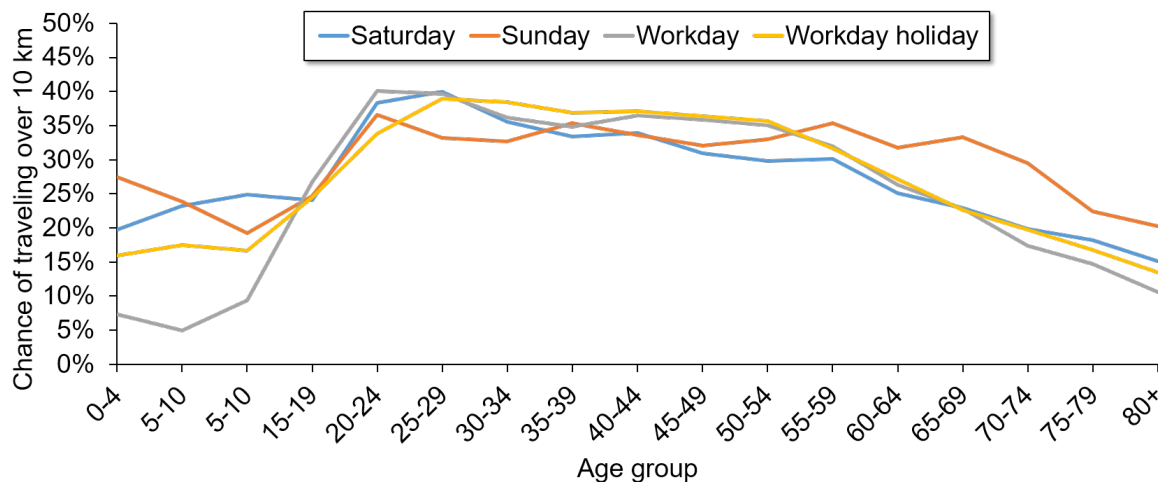


FIGURE 5.10: The chance of making a trip greater than 10 kilometres per age group. This data is extracted from [OViN](#).

we are able to retain a proper sample size per group to allows us to add more dimensions, which we will discuss in the next paragraph. Ages greater than 80 are grouped to achieve a meaningful sample size for the elderly age groups.

Next to age various other dimensions have been explored to find dimensions in which variation is present, while having sample sizes that are to a lesser degree sensitive to outliers. In this case the notion of sample size receives additional importance, because when groups have a small sample size outliers have a great impact on the mean and will therefore cause lots of variation. However, variation caused by outliers is not the variation we are looking for. Therefore, we must not look at the variation alone but also validate that the sample size per group is sufficient in order for outliers to make less of a difference on the mean of the group. We used an arbitrary minimum sample size of 20 per group before any variation was considered to be "true" variation. Using this criterion we found, next to age, variation in three other dimensions, namely: workdays, Saturdays, Sundays and workdays during holidays. There might also be some difference between workdays, but—according to our own criterion—we do not have enough data to properly prove these distinctions.

For the purpose of our research we propose a new scaling method that takes two limitations of the current scaling method, discussed in the previous two paragraphs, into account. The proposed scaling method is visualised in figure 5.11. On the left hand side of this figure are the processes of the proposed scaling method. Depicted on the right hand side of this figure are the deliverables that are the result of executing the corresponding processes. The result of

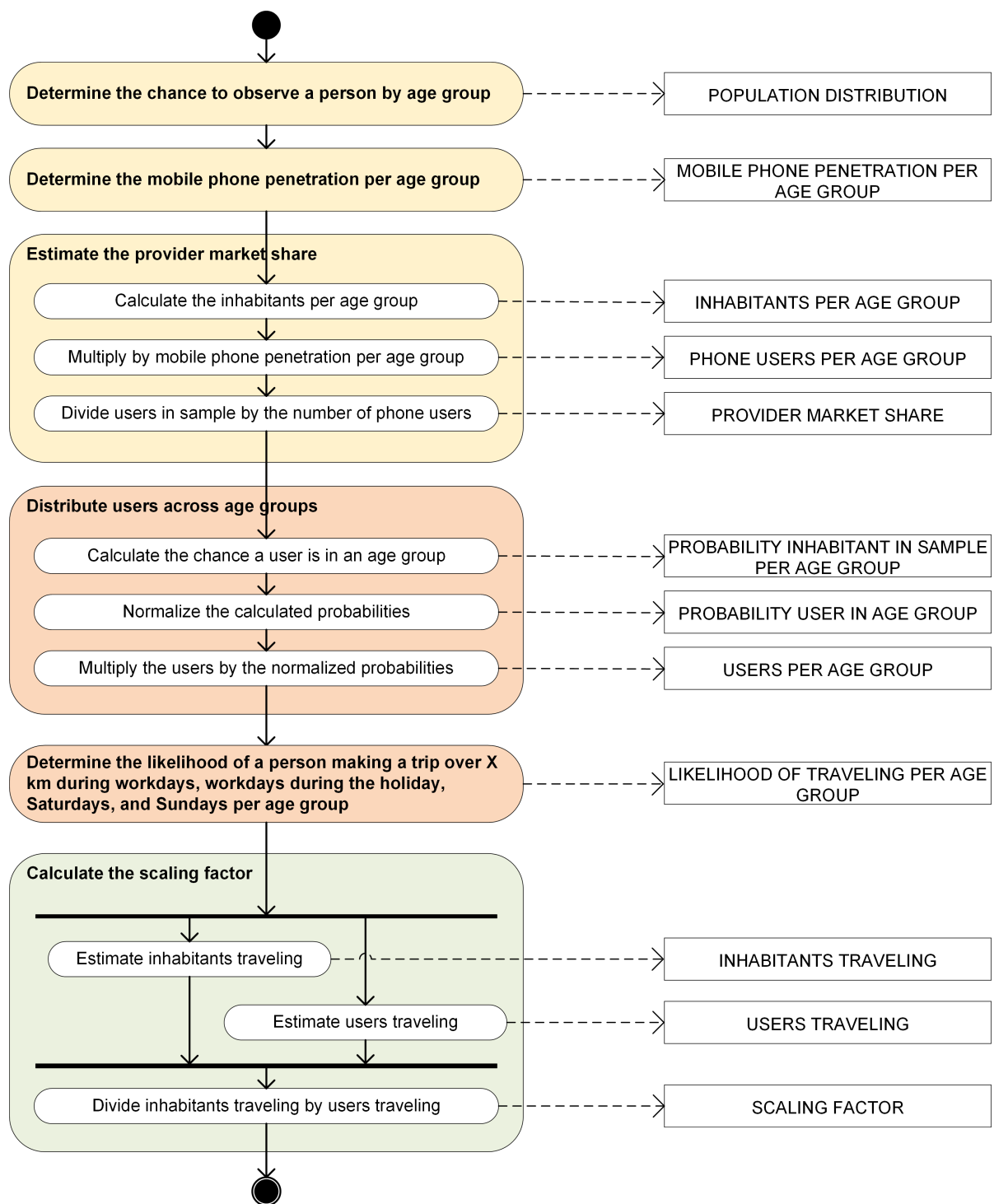


FIGURE 5.11: The improved method to determine the scaling factor, which can be used to extrapolate the sample to the travelling population.

performing the entire method is a scaling factor that applies to a specific Mezuro area for a specific day. Thus, to compute the proposed scaling factor for the entire dataset for a single month this method needs to be performed 37.770 times (which is equal to the 1259 Mezuro areas times the number of days per month).

Detailed descriptions of the concepts and activities in figure 5.11 are provided in the tables in appendix G. Table 5.4 gives an example of the how the scaling factor is calculated, which we explain in the following paragraph.

TABLE 5.4: A simplified and fictional example of the calculation of the OD scaling factor for a single Mezuro area for a single day. This example only includes the dimension age group (grouped by 20 years). The full method also includes the distinction between Saturdays, Sundays, normal workdays (Monday to Friday) and workdays during holidays but these are omitted in this example to the improve readability.

	Groups					Total	Calculation	Source
	1	2	3	4	5			
Constants								
A	Active users	4000						Mezuro
B	Present inhabitants	15000						CBS
C	Age groups	0-19	20-39	40-59	60-79	80+		Our research
Market share								
D	% Age group in area	15%	40%	15%	20%	10%	100%	CBS
E	% With mobile phone	50%	90%	84%	62%	19%	-	TelecomPaper
F	Inhabitants in area	2250	6000	2250	3000	1500	15000	B*D
G	Inhabitants with phone	1125	5400	1890	1860	285	10560	E*F
H	% Market share	38%	38%	38%	38%	38%	-	A/SUM(G)
Users traveling								
I	Multiplier	2,8%	13,6%	4,8%	4,7%	0,7%	-	D*E*H
J	Normalized multiplier	10,7%	51,1%	17,9%	17,6%	2,7%	100%	I/SUM(I)
K	Users	426	2045	716	705	108	4000	A*J
L	Chance of trip >10 km	5%	50%	40%	30%	10%	-	OViN
Scaling factor								
M	Inhabitants traveling	112,5	3000	900	900	150	5062,5	F*L
N	Users traveling	21,3	1022,7	286,4	211,4	10,8	1552,6	K*L
O	Scaling factor OD	3,26						SUM(M)/SUM(N)

The calculation of the scaling factor in the example is performed in the same sequence as the method. Additionally, the colour coding can be used to cross reference the example to the method. We now elaborate the calculation of the scaling factor in table 5.4. First, the distribution of the population in the Mezuro area over the age groups is determined (row D). Second, the mobile phone penetration per age group is determined (row E). Third, the market share is calculated in three sub steps. The inhabitants in the Mezuro area are distributed according to the population distribution (row F). Then the number of inhabitants with a mobile phone is determined (row G). Subsequently, the market share of the telecom provider in the selected Mezuro area is calculated by dividing the number of active users by the sum of all present inhabitants with a mobile phone (row G).

The next step is to distribute the active users across the age groups. We obtain the multiplier by multiplying the chances per age group (row I). The multiplier represents the chance that a random person from the present population of the Mezero area is in our sample, i.e., an active user. We normalise this multiplier (row J). The normalised multiplier represents the chance that a random person from our sample belongs to that age group. The next step is to multiply the normalised multiplier by the number of users. This results in the distribution of active users over the age groups (row K). The next is to determine the likelihood that someone of a certain age group makes a trip greater than 10 kilometres (row L), which is the data depicted in figure 5.10.

The last three steps are used to determine the travelling users, the travelling population and the ratio between them, i.e., the scaling factor. The travelling population is determined by multiplying the inhabitants per age group by the likelihood that someone of a certain age group makes a trip greater than 10 kilometres (row M). The number of users travelling per age group is calculated by multiplying the active users per age group by the likelihood that someone of a certain age group makes a trip greater than 10 kilometres (row N). Finally, the scaling factor is calculated by dividing the number of inhabitants travelling by the number of users travelling (row O). Note that this example given does not include the dimensions Saturdays, Sundays, normal workdays (Monday to Friday) and workdays during holidays in order to improve the readability of the example. However, the method as formalised in our PDD (see figure 5.11) does include these dimensions.

Due to limited availability and reliability of data from market research on the market share of the provider per age group we had to take another approach. This made us choose to calculate the market share based on the ratio between the active users and the total population with a mobile phone. The underlying assumption here is that everyone with a mobile phone has a provider. By calculating the market share this way, regional differences in the provider market share are compensated for. However, this means that from a mathematical point of view the inclusion of the market share in the calculation of the scaling factor has merely become irrelevant. That is because the resulting scaling factor only reflects differences in proportions between age groups and by including the market share as an equal value for all age groups there are no differences between age groups. Hence, fluctuations in the market share no longer affect the resulting scaling factor. Although differences in the market share per area are not reflected in a change of the scaling factor, we like to keep it in the method for the purpose of completeness. Moreover,

when in the future more data becomes available on the market share per age group this can easily be added to the method.

5.7 Evaluating the Improved Scaling Method

Now that there is an improved method to scale the sample to the travelling population it needs to be applied and the results need to be evaluated. Applying the scaling factor is a pretty straightforward process and does not need much explanation other than that we simply substitute the original scaling factors for the improved scaling factors. Therefore, in this section we will focus on the evaluating whether the number of people travelling inferred from the mobile phone location data corresponds to another independent data source that can provide these numbers.

Finding such a data source is no easy task. At first glance [OVIN](#) might seem like an option. However, one of the main drawbacks of [OVIN](#) is that the sample size is not large enough to use it for counting people on specific locations. Luckily, there is an alternative available that does provide a sufficient sample size to make this comparison, which is the road side measurements data. This data source is completely independent from the mobile phone location data as it is not used in the process to determine the travelling population based on the mobile phone location data and therefore an ideal data source for us to use as comparison.

The road side measurement data is distributed by the Nationale Databank Wegverkeergegevens (NDW) and freely available. The [NDW](#) is a public organisation that collects, stores, and distributes various types of measurement data related to road traffic. Their goal is to support and improve traffic management. Most of the data they have is obtained through the inductive-loop vehicle detector, which is integrated in the surface of the Dutch road network. An example of this type of this detector is depicted in [figure 5.12](#). The self-reported accuracy of these types of measurement devices is over 90% with a mean of 99% ([Meppelink, 2016](#)). Although the data is openly available through the web, its verbosity makes it difficult to process. Nonetheless, [Meppelink \(2016\)](#) managed to extract the vehicle counts from the road side measurement data for his research and compared it to the travelling population inferred from the mobile phone location data.

The advantage of using this data for the evaluation is that it is relative easily available, representative at a local scale, provides accurate results and available at a minute-based interval.



FIGURE 5.12: This inductive loop detector system is ubiquitous in the Dutch road network. The vehicle counts obtained through this system can be used to evaluate the travelling population inferred from the mobile phone location data.

Since we lack the data to compare the mobile phone location data for other modalities than the vehicle counts on the roads, we assume that the results of this comparison can be generalised to public transport and other modalities. [Meppelink \(2016\)](#) investigated whether (1) there is a correlation between travelling population from the mobile phone location data and the road side measurements and (2) if the absolute number of vehicles match the number of people travelling inferred from the mobile phone location data. Note that the [NDW](#) data provides vehicle counts, which is different from the people count inferred from the mobile phone location data. A good correlation confirms that the patterns match in relative numbers. Additionally, a good match in absolute values confirms that the scaling factor is calculated and applied properly.

In order to go from people to vehicles a correction is applied to the travelling population equal to the average number of people per car, which is described in detail in [Meppelink \(2016\)](#). In his analysis he compared road side measurement sites spread evenly across the Netherlands to the mobile phone location data. He found that the correlation between the two data sets ranges from .92 up to .98, depending on the situation. He concludes that there is a strong indication that the patterns of the road traffic can be very well inferred using mobile phone location data.

In his comparison on the absolute number of vehicles he found that the two data sources are a very good match on some sites, while others show some deviation. However, overall the measured vehicle counts and the inferred vehicle counts are a good match leading us to believe that the improved scaling method provides results that can be used in the remainder of this research.

5.8 Conclusions

The negative effects observed by looking at descriptive statistics of the mobile phone location data are irrelevant as they occur outside the period that is analysed within the context of this research. The positive effects observed however, are influential as the increased adoption of 4G technology by the majority of the subscribers leads to more events being produced, which in the long run leads to more accurate location estimation.

Trips under 10 kilometres are not accurately detected using the mobile phone location data. Therefore, these trips are omitted from this research. Next to that, we found that cells with cell radii larger than 12.5 kilometres actually deteriorate the current location estimation algorithm and are therefore omitted in this research. This causes 4% of the events to be dismissed, as these events are handled by cells with radii larger than 12.5 kilometres, but improves the location estimation.

Next to that, three measures are taken to satisfy privacy regulations. First, data can only be processed at the mobile phone service provider. Second, mobile phone numbers are hashed using a hash key that changes every month. Third, outputs are aggregated and can only be viewed when they contain 15 users or more.

Furthermore, the current method used by Mezero to go from devices to the travelling population does not correct for any of the demographic differences between the Dutch travelling population and the subscribers in the sample. We proposed a new method that takes into account the differences per age group in the penetration of mobile phones and the likelihood of performing a trip greater than 10 kilometres. The results have been compared to road side measurements data and the overall picture of all trips within the Netherlands corresponds to a great extent confirming that the proposed scaling method delivers acceptable results.

Chapter 6

Trip Motive Prediction

It is very important to understand that the attractiveness of a city differs for various groups within our society (Sinkienė & Kromalcaš, 2010). For example, when someone is interested in buying new clothes he will look for department store. On the other hand, a person that travels to a city to visit his family generally is not interested in the number of department stores. Hence, the relative importance of the factors that influence a person's decision to visit a particular city differ per motive for visiting. Therefore, in this chapter, our goal is to provide an answer to **SRQ3**: *How can motives for visiting a city be identified from mobile phone location data?*

For this a model has to be created that can determine trip motives based on a set of trip and person characteristics. To create the model the **CRISP-DM** method is applied as discussed in the research approach. The following six sections will contain the products resulting from performing the **CRISP-DM** method. Ultimately, this will lead to a model that can be used to determine and add trip motives to the mobile phone location data.

6.1 Business Understanding

It is important to understand that we always look at travel behaviour from a population or aggregated perspective. The fact that trip motives on the level of the individual are correct are of lesser importance because we are interested in the behaviour of the masses. Moreover, in accordance with the minimum rule of 15 we are not able to observe visitor flows that aggregate less than 15 people. Hence, the aim is more geared towards predicting how the motives are distributed for all trips rather than predicting well on the individual level.

In the end the goal is to use the model to make predictions on the mobile phone location data. Because of technical constructs there are a few crucial points to take into account when designing the model. First of all, for this research, one can only interact with the mobile phone location data using [SQL](#) queries. This is important to keep in mind because some models are easier to convert to [SQL](#) than others. Moreover, to be of practical use in the long run the query has to run faster than it takes new data to enter the database. For example, if it takes two days to process one day of data this is unsustainable. Ideally it takes much less than a day (e.g., one hour max) to also be able to add trip motives to historical data and allow for other processes to run.

In summary there are three critical success factors. First, the model needs to significantly improve the baseline estimation, i.e. the a-priori information of the underlying distributions, on the classes home-to-work, business, services, shopping, education, visiting, other social recreational, touring, and other. Second, the model needs to be applicable in an [SQL](#) type of environment. Third, implemented model should be able to determine the trip motives of all trips in the mobile phone location data within a reasonable time frame, e.g. approximately one hour.

6.2 Data Understanding

To link trip and or people characteristics to trip motive it is key to have a data source in which both are present. In the Netherlands the largest source of information containing both trip characteristics and trip motives is the [OVIN](#). This is a yearly survey aimed to gain information about mobility patterns of inhabitants of the Netherlands. The survey is performed by [CBS](#), i.e. a large organisation that is tasked to gather and present statistics about the country and its inhabitants. Other data sources about mobility patterns in the Netherlands do exist. The most noteworthy is from the Mobiliteitspanel Nederland (MPN). The key differences are that the [MPN](#) covers three days in a person's life and [OVIN](#) just one, but at the cost of a ten times smaller sample size. Moreover, the three days available appear to be only a slight addition. Weekly patters, for example, are still not visible with the [MPN](#). Because of the superior sample size and the fact [MPN](#) still does not allow to see weekly patters, the [OVIN](#) is preferred over the [MPN](#). This does imply we can only distinguish daily patterns and cannot identify travel patterns that span multiple days, e.g. a person always going to work at the same location. Additional information could potentially be generated by employing both data sources. Combining these data sources would require significant effort and given the limited time available for this research

this will be left for future research. Moreover, if the model created just using OViN produces satisfactory results there may be no incentive to put in the extra effort.

For this research the nine trip motives that are distinguished within OViN are used. These are home-to-work, business, services, shopping, education, visiting, other social recreational, touring, and other. We will provide the definitions for these according to OViN.

Home to work

Trips to and from a work location. This can be a regular as well as temporary work location. Hence, this also includes locations for on-call employees, part-timers and voluntary workers (CBS, 2014).

Business

Business trips are trips that are due to work, excluding trips that are to regular work locations. This category mainly concerns service trips, customer visits, meetings and symposiums (CBS, 2014).

Services

These trips are towards a location where someone can get a service or can get personal care. Examples include visits to the city hall, the barber, the general practitioner or a mortgage advisor (CBS, 2014).

Shopping

Concerning the purchase of goods. These purchases can be food, such as groceries, as well as non-food. The purchase itself is not a requirement. Therefore, browsing stores or windowshopping is also included in this motive (CBS, 2014).

Education

Concerns all that has to do with education and childcare (CBS, 2014).

Visiting

Concerns the visiting and possibly staying over at friends, family or relatives (CBS, 2014).

Social recreational other

Includes trips for the purpose of sports or trips to places to perform sports, hobbies or watch them. It also includes visits to the hospitality sector, cultural activities and religious activities (CBS, 2014).

Touring

These are trips that have the same origin and destination, like taking the dog for a walk. However, it is not limited to walking and also includes other modes of transportation such as bikes and cars. Explicitly, these trips do not have the destination as goal. Rather enjoying the trip itself is the goal (CBS, 2014).

Other

Includes the non-commercial activity of picking up and delivering of persons or goods. Also includes the travelling to places such as a parking, train station or car. Furthermore, trips that do not belong to any other category belong to this category (CBS, 2014).

For this study the results of the OViN surveys for the years 2013 and 2014 are combined (CBS, 2014). The results found in these OViN surveys has been analysed by the CBS and both datasets passed the tests (CBS, 2015). This implies that the data appears logical on a high level. The surveys are similar in the structure and size between the two years. In 2013 and 2014 OViN has 34,452 and 34,826 respondents, respectively. Weight factors are included in the results to compensate for biases in the sample. These weight factors can compensate for biases on household, person, and trip level. Because we try to predict motives for trips the weight factors on trip levels need to be included when training our model. In total the dataset contains information on 254,317 trips spread over the two years.

Because of the earlier design decision to focus on trips longer than 10 km, only a subset of the results from OViN are relevant. OViN does have distance classes to state the distance of a trip, but these are not necessarily as the crow flies. In the mobile phone location data only trips over 10 km are taken where the distance is measured as the crow flies. Fortunately, the OViN does include the four digit postal codes of the origin and destination. An additional database retrieved from postcodedata.nl is merged with OViN to add GPS coordinates to the origin and destination (Postcode Data, 2014). Before adding the GPS coordinates to the dataset the GPS coordinates per postal code, e.g. 1234AA, are averaged to get the average for the four digit postal code as present in OViN. In data preparation the GPS coordinates are translated to distances and by taking into account the reported travel times also the velocities are calculated.

When performing a quick scan over the velocities calculated we encounter some strange behaviour. Occasionally the velocities for trips over 10 km are much greater than can be physically possible and are incoherent with other answers. This is most likely because people might not

know the postal code they are in or travelling towards and make mistakes here. Characteristics of the origin and destination may provide insights into why people travel. Therefore, it is important to ensure the final dataset includes only trips where it is at least probable that a person travelled from origin to the destination. Improbable trips will be discarded. These are trips with a velocity greater than 145 km/h. The 145 km/h threshold is chosen based on the 130 km/h maximum velocity that is allowed on Dutch highways. Because some people might go beyond the maximum allowed velocity an additional 15 km/h are added. Note that the distance is as the crow flies and not over the road and hence the 145 km/h constraint is a bit more lenient in practice. Only the extreme outliers will be left out. How this affects the dataset size will be further discussed in the data preparation section.

The mobile phone location data is also an important dataset that needs to be understood here. On this dataset the model has to be applied. Hence, the characteristics used to predict trip motive also have to be present in this dataset. In table 6.1 a list of the fundamental characteristics in the mobile phone location data is presented along with a short description.

For the arrival and departure time there is an important distinction between [OVIN](#) and the mobile phone location data. [OVIN](#) contains information about when a trip starts and the mobile phone location data about events at the destination and origin. In the mobile phone location data the last event before departure is used as the actual departure time, which may result in incorrect arrival and departure times. It may be the case that a person has his last event at 10 o'clock and actually leaves at 11 o'clock. On the other hand, it may be that a person leaves at 10 o'clock, but only gets outside of the range of the cell at the origin such that the

TABLE 6.1: Trip attributes present in the mobile phone location data.

Attribute	Description
Origin/Destination	Area of origin / destination of a trip. The area may be a municipality or sub-part of a municipality. For densely populated areas, e.g. Amsterdam, there are several areas in one municipality.
Time of departure	The last CDR inside the range of the cell tower at the origin.
Time of arrival	The first CDR inside the range of the cell tower a person at the destination.
Time at destination	The total time after a trip before the next trip starts.
Home based	This attribute states if the person is either departing home, leaving home, or travelling between non homebased places. Home is where a person spends most evenings (between 8pm and 7am) during weekdays and weekends (Geerts, 2014). This is calculated per person on a monthly basis and is specified to a level of detail of a four digit postal code, which is smaller than a municipality or an area.

last event can still be at a few minutes past 10 o'clock. This is important to keep in mind when implementing the model.

The homebased attribute, which indicates if a trip originates from or arrives at home, is present in the mobile phone location data. However, by default this attribute is not present in the [OVIN](#). We expect homebased to be an important attribute to predict, for example, the home-to-work travel motive. Therefore, this attribute is constructed for [OVIN](#). How the homebased attribute is added will be discussed in the following section, i.e. data preparation.

6.3 Data Preparation

In this section we will discuss how additional attributes are constructed, and how the dataset is filtered to contain only useful information.

All attributes that are used in the model to predict trip motive are presented below this paragraph. A description is provided stating exactly how these attributes have been created. The order of presentation is also the order in which the attributes are created. Some attributes are created before the dataset is reduced because of quality concerns stated in the section '[Data Understanding](#)' or because it is irrelevant to our research, e.g. information outside the scope of this research. The order in which the steps are presented is important because it affects the calculation of some attributes. For example, by only including trips with a travel distance of at least 10 km it influences how many trips a person has taken on a day. Therefore, the filter on the travel distance of trips needs to be applied before the number of trips per person is calculated. The name of each attribute is shown in **bold** and when data reduction is performed this is stated in *italic*. A description of each attribute can be found at the end of this section in table [6.2](#).

The first data cleansing step is removing duplicates in [OVIN](#). In [OVIN](#) each trip can also have sub trips. For example, a trip includes going to the train station by foot, taking the train, and going to work by foot. Then the motive is assigned to all three trips. And the main means of transportation, i.e. the train, is also noted three times. As we are only interested in the main means of transportation we would get three rather than the correct one trip recorded. Removing 'duplicates' reduces the number of trips from 271,824 to 248,807.

Homebased

In [OVIN](#) there is an attribute stating the goal of a trip. If this goal is going home we know the postal code of the destination is the home postal code. For all people that once stated the goal of their trip is going home we can thus infer the home location. Based on this we can decide on a person level if someone is leaving home, going home, or is travelling between locations that do not include the home postal code.

Travel distance

Travel distance is calculated by linking the postal codes of the origin and destination location with their corresponding longitude and latitude, i.e. [GPS](#) coordinates. The link between postal codes, longitude, and latitude are provided by an external data source ([Postcode Data, 2014](#)). Originally the external data source includes postal codes with digits and two characters, e.g. 1234AA. Longitude and latitude are averaged per four digit postal code, e.g. 1234. The longitude and latitude are then linked to the origin postal code and destination postal code. Using the R package [Geosphere](#) and specifically with the function `distm()` the travel distance is calculated from these longitude and latitude combinations ([Hijmans, 2015](#)).

All trips under 10 km are discarded as they fall outside of the scope of this research. This implies the dataset is reduced from 248,807 to 62,723 trips (23.1%).

Trip start / end

Values are calculated by multiplying the start / end hour of a trip by 60 minutes and adding the start / end minute of a trip. Start / end hour and start / end minute are readily available in [OVIN](#).

First / last trip start / end

The start / end of the first trip is calculated by aggregating over the entire [OVIN](#) dataset per person id and taking the start / end time of the first trip. The first trip start / end is added to all trips taken for each person. The same is also done for the last trip of the day on a person level.

Velocity

The velocity of each trip is calculated by dividing the trip's travel distance (converted from metres to kilometres) by the difference between trip's end time and trip's start time (converted from minutes and hours to minutes).

TABLE 6.2: Trip attributes used in modelling and evaluation.

Attribute	Description
Trip motive	The motive of a trip. In this research nine different motives are distinguished, which are named earlier in this chapter.
Arrival / Departure time	This is the start and end time of a trip. Trip start and end are both expressed in minutes from 0:00.
First trip start / end	Start and end of the first trip for a person. This attribute is included because, for example, people starting the day early might be more likely to have business trips later in the day.
Last trip start / end	Same as the previous. When a person ends a day of travelling might indicate what type of traveller a person is and so help predict trip motives of other trips during the day.
Homebased	This attribute states if a person is leaving home, arriving home, or has a trip not involving his/her home location.
Holiday	Whether or not it is a holiday at the persons home location.

As stated in data understanding, section 6.2, the trips over 145 km/h are unrealistic and either the origin or destination is incorrectly reported. These trips are, therefore, also left out. This results in a data reduction of 5,012 trips. Moreover, although this may have been done earlier, the trips that are exclusively abroad are omitted. This is just a reduction of a mere 2 trips. Finally, trips with no value for the attribute trip motive that we aim to predict are removed. These add up to a total of 17,373 trips. The final dataset used for modelling and evaluation consists of 40,522 trips.

Weekday

Day of the week might also provide insight into how people travel. This characteristic might even be one of the more important ones. The reason that it is at the end of the data preparation stage is that it requires less time to do the calculation here, i.e. after irrelevant trips have been performed. Weekday is a simple conversion from the date variable that is already present in **OVIN** and added as numerical values such that Sunday up to Saturday get assigned the values 0 up to 6, respectively.

Holiday

During holidays the distribution of trips changes significantly as the number of work related trips decreases. Therefore, each trip is assigned whether it happens during a holiday or not.

6.4 Modelling

The modelling phase consists of two distinct steps. In the first step an algorithm to train the model is chosen. In the second step we determine how to use the chosen algorithm in order to get the best model as a result. Subsection 6.4.1 will discuss what type of algorithm fits our task best. Subsection 6.4.2, thereafter, focuses on how to apply the algorithm, i.e. what settings should be used to obtain the best possible model for our purpose.

6.4.1 Model Choice

For model choice it is important to take into account the structure of the available data and our goals, which are elaborated upon in the business understanding phase, discussed in section 6.1. What is apparent from the data is that we have nine distinct classes of trip motives to predict. The model chosen will thus have to be a classification rather than regression type of model. Furthermore, there are approximately twelve input variables and all of these are either numeric or binary (0, 1). There is one exception, i.e. the homebased variable. The homebased variable can have the value -1 when the home location is the origin, 1 if the home location is the destination and 0 when the home location is not involved in the trip. However, if required the homebased variable can be transformed by creating two binary variables. To our knowledge all popular classification algorithms allow for numeric data and thus this does not provide any noteworthy restrictions on our model choice. On the business side it is important that the model can be implemented relatively easily into [SQL](#) and that the implementation is relatively fast, e.g. runs a day of data within one hour. These constraints, as discussed earlier, are crucial for the model to be useful in practice. Moreover, these constraints significantly narrow the number of candidate classification algorithms down.

Neural networks and support vector machines have shown to provide solid results ([Caruana & Niculescu-Mizil, 2006](#)). However, implementing these algorithms in [SQL](#) is no trivial task. Neural networks and support vector machines also have the disadvantages of being more difficult to explain and visualise. Decision trees and [PETs](#) are much easier to comprehend and convert into [SQL](#). The trees are basically an ordered set of if-else-statements that will lead to a value. In case of decision trees the values are hard label, e.g. business trip, and for [PETs](#) the values contain probabilities, e.g. services (1% chance), work (11% chance), education (1% chance), touring (2% chance), recreation (28% chance), visit (29% chance), shopping (19% chance),

business (1% chance), other (8% chance). For predictions on groups rather than individuals soft labels, i.e. probabilities, are generally preferred (Provost & Domingos, 2000). In the above example, if 100 people would have been labelled with a hard label we would get 100 trips for the purpose of visiting friends or family. With a soft label it would be 1 trip for services, 11 trips to or from work, 1 trip for the purpose of education, etcetera. The latter will typically be a better representation of the underlying population. Hence, **PETs** are preferred for our model over the standard decision tree.

Another well-known classification algorithm is decision forest, which is basically a large collection of small decision trees. Each tree in the forest provides a prediction and based on all the predictions one final verdict is given. This could also be in the shape of a hard and soft label, i.e. probabilities of belonging to a class. Because a decision forest is in essence a collection of decision trees it is also a collection of if-else-statements. However, a decision forest would take much longer to make predictions. For example, a large decision tree with a tree depth of 20 is already large, but would per run at most take 20 if-else-statements. A decision forest works well when there are hundreds or thousands of small trees. If each tree would only consist of one if-else-statement it might require orders of magnitude more in computation time.

For this study we will implement a **PET**. The **PET** is preferred over the alternatives because it is (1) easy to translate into **SQL**, (2) provides a better estimation of the distribution than a decision tree, and (3) requires little computation time at implementation compared to other alternatives. To construct the **PET** the `rpart` package for R is used (Therneau, Atkinson, & Ripley, 2014).

6.4.2 Training a **PET**

Numerous studies have been performed to determine how to best predict class probabilities using **PETs** (see for example Ferri, Flach, & Hernández-Orallo, 2003; Margineantu & Dietterich, 2003; Richards et al., 2012; Rüping, 2006; Zadrozny & Elkan, 2002). Because the goal of decision trees, i.e. predicting hard labels with maximum accuracy, differs from **PETs**, i.e. estimating class probabilities, the two models have to be trained and evaluated differently (Provost & Domingos, 2003). For decision trees the standard procedure is to train the tree and perform pruning afterwards (Provost & Domingos, 2003). Pruning is performed to check if leaves are worth the added complexity they bring to the model and remove them if they are not (Provost & Domingos, 2000; Zadrozny & Elkan, 2002). The idea is that complex trees model are likely to

over fit the training data. Over fitting implies that the model incorporates too much noise and outliers and thus reduces the predictive power of the model on unseen data. Predicting well on unseen data is obviously the goal and, therefore, pruning is crucial.

For probability estimation trees, however, the story is a little different (Provost & Domingos, 2000; Zadrozny & Elkan, 2002). For getting good estimates pruning can also be a culprit, hurting results (Provost & Domingos, 2000; Zadrozny & Elkan, 2002). While pruning removes outliers it also tends to remove leaves that have little predictive power because the underlying probabilities are too similar. The latter would actually help the effectiveness of probability estimation trees, because these focus on distributions and changes herein rather than the best accuracy of predicting classes (Provost & Domingos, 2000). An example of this can be found in figure 6.1. The bottom left two leaves do not truly add to the predictive power. In both cases the tree would predict just as many good or wrong as when these leaves would be pruned. However, for a PET the bottom two leaves do add value. They show the chance on the bottom right leaf is 50/50 and that of the left is 25/75, i.e. a noteworthy difference.

Provost and Domingos (2003) tested how well probability estimation trees would perform with a variety of settings. They did this on 25 publicly available dataset commonly used for testing data mining algorithms, e.g. the Iris and Hepatitis datasets that are also built into R (Provost & Domingos, 2003). Pruning, for one, performed better than not pruning. They hypothesise that even though pruning removes useful information it also removes results in leaves that are produced from very few observations (Provost & Domingos, 2003). It may, for example, be the case that a distribution at a leaf node is only built on five observations that all belong to the same class. As a result this leaf will assign all future observations that end at that leaf with a 100% certainty to that class. Of course this can be correct, however, with a small sample size it is likely to be incorrect.

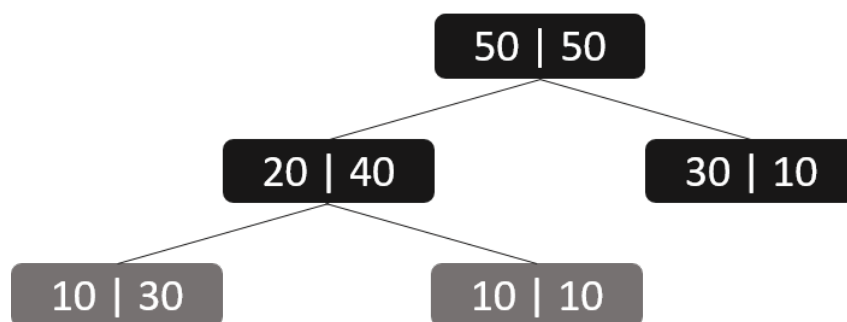


FIGURE 6.1: An example of a PET.

Laplace correction is one technique occasionally implemented to address the issue of overly confident distributions at leaf nodes with few observations. Laplace correction is applied to reduce the confidence of the leaf by adding one observation to each of the classes artificially. Consequently, instead of having five observations of one class you get six of that class and one of each other class. This will result in a more uniform distribution. When applying Laplace correction on the unpruned tree the results found were slightly better than with pruning, although not significant (Provost & Domingos, 2003).

Zadrozny and Elkan (2002), for one, point out that applying Laplace correction might not be ideal as it results in a more uniform distribution, which might not be the true underlying distribution. Rather than using Laplace correction they went for m-estimation in their study (Zadrozny & Elkan, 2002). M-estimation is largely similar to the Laplace correction, but draws the distribution closer towards the a-priori distribution rather than a uniform distributions.

Equations 6.1 and 6.2 will show how class probability is calculated before and after m-estimation, respectively. In the equations p_i stands for the probability of class i , k_i stands for the number of observations of class i at the leaf, and n stands for the total number of observations of the leaf of interest. b_i is the a-priori probability of encountering class i and m is a multiplier that determines how much smoothing is applied. The optimal value for m can be determined by using cross validation (Cussens, 1993).

$$p_i = \frac{k_i}{n} \quad (6.1)$$

$$p_i = \frac{k_i + m * b_i}{n + m} \quad (6.2)$$

An alternative to smoothing, e.g. by applying m-estimation, is to perform isotonic regression (Niculescu-Mizil & Caruana, 2005). The issue with isotonic regression is that it is hard to implement in our situation. Isotonic regression implies the class probability has to be continuously increasing or decreasing and that is not necessarily the case in our situation. Moreover, it is designed to be used for binary classification and extending the algorithm to a multi-class problem is non-trivial (Niculescu-Mizil & Caruana, 2005).

In addition to applying smoothing, [Zadrozny and Elkan \(2001\)](#) propose to perform curtailment. Curtailment is the act of ignoring leaves that have less than to be established number of observations ([Zadrozny & Elkan, 2001](#)). This differs from traditional methods to reduce the complexity of a tree such as applying a threshold on the number of observations each leaf should have during creation of the tree. Curtailment occurs after a tree is created.

In figure 6.2 an example of a tree is provided where the grey leaf is a leaf that is being ignored because it does not meet the minimum number of observations criteria. When an observation would end up in a grey leaf the distribution of the parent node will be used, provided it has sufficient observations otherwise it will go to the grandparent. This process can continue all the way back to the root node. Even though [Zadrozny and Elkan \(2001\)](#) show promising results when applying curtailment we show here that curtailment can result in biases in prediction. Imagine each leaf should have at least 20 observations to provide sound estimations of the distribution. Curtailment will then result in the black tree as shown in figure 6.2 where leaf 5 is left out because the lack of observations. Whenever an observation ends up in that leaf during prediction the distribution of node 2, i.e. its parent node, will be assigned. Now imagine 60 new observations entering node 2 and we want to predict the distribution. If the tree is correct this would result in 20 yesses and 40 no's, i.e. using the left class as yes and the right class as no. In the tree after curtailment about forty-five observations go to node 4, which results in 10 yesses, and fifteen go to node 2 as node 5 does not meet the criteria, which results in a further 5 yesses. In total we get 15 rather than the expected and correct 20 yesses. This example shows curtailment can result in artificial biases and, therefore, we abstain from using it. The general idea of setting a minimum to the number of observations that should be in a leaf appears useful nonetheless. For this we propose to use `minleaf`, which is an old technique that only allows the

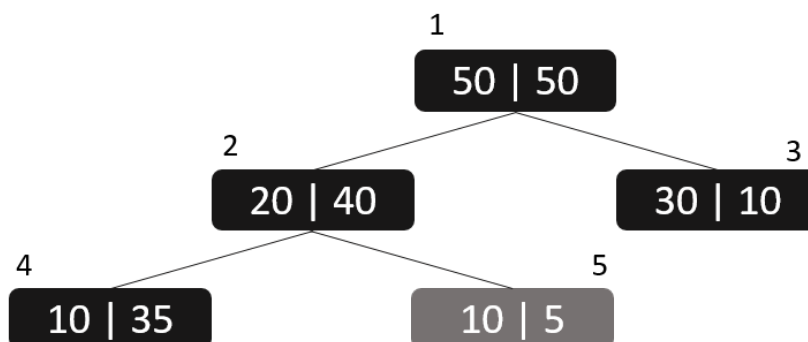


FIGURE 6.2: An example of a [PET](#).

tree to grow further leaves if at least a certain number of observations will end up in each leaf. Setting `minleaf` at 20 in our example would have resulted in a tree without nodes 4 and 5 which implies the 20/40 distribution at leaf 2 would be assigned to new observations, which is correct. Cross validation, as with curtailment, will be used to determine what the threshold on `minleaf` should be to deliver the best results.

In summary, **PETs** are different from decision trees in terms of function and hence should also be trained differently (Provost & Domingos, 2003). Traditional pruning that is often performed when creating a, or actually after creating an initial, decision tree will most likely not lead to the best results for **PETs** and we advise not to apply it. By performing pruning valuable information in the underlying distribution that might not help decision trees is removed even though it could still be valuable for **PETs**. Without pruning there is a very good chance the tree will over fit the training data and by definition not work well on unseen data. Alternatives to pruning need to be applied that prevent **PETs** from making strong claims without having sufficient data to back these claims up, e.g. a leaf with two observations stating all new observations belong with 100% likelihood to class X. One technique to prevent overly confident estimations is smoothing. M-estimation, which is a specific smoothing technique, seems a prime candidate and works by adding dummy variables following the a-priori distribution to each leaf, after which the likelihoods are recalculated. This will cause the distributions in each leaf to be less extreme and closer to the a-priori distribution. Alternatively, or in addition to smoothing, one could also stop growing leaves with few observations. This can be done by only allowing splits in the tree where each leaf node needs to have at least a certain number of observations. When training a **PET** the number of dummy variables added with m-estimation and minimum number of observations needed can be adjusted. Evaluation of the created models under different settings will determine the best settings for a particular situation.

6.5 Evaluation

Evaluation is about finding out how good our model performs under different circumstances when trained using a variety of settings. Finally, this will help us to answer questions such as “Does the model add to what we already know?”, “How accurate is the model in different circumstances?”, and “What settings will result in the best model?”.

This section is divided in four subsections. In 6.5.1 the method to correctly evaluate the model is described. In 6.5.2 the evaluation results are presented with our conclusions following in the subsequent section, i.e. section 6.5.3. A description of the best model including information about what attributes add much to the predictive power of the model can be found in section 6.5.4.

6.5.1 Evaluation Method

When evaluating the created models it is key to have a good measure of how good a model is. Moreover, the measure will have to fit the task the model is designed for. A popular quote states: “Everybody is a genius. But if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid.” The same goes for model evaluation. The goal of our model is to provide accurate estimations of the distribution of trip motives. We do not necessarily care about each individual. Hence, the goodness measure should be about predicting distributions accurately and not individual observations. As a goodness measure we will use the Chi-square test. The Chi-square test evaluates whether a distribution, in our case the estimated distribution, might represent the actual distribution, i.e. the distribution in the test set. When the distributions do not significantly differ, with a confidence level of .95 (α is .05), we assume the prediction is correct and otherwise that the prediction is false.

The Chi-square test has two assumptions that have to be satisfied for the test to be meaningful. These are:

- Independence of observations, i.e. each observation needs to be unique and unrelated to any other observation (Field, 2009).
- Each class should have at least five observations. When there are less than five observations it is generally assumed the test has too little statistical power (Field, 2009). Hence, with less than five observations per class the test does not provide conclusive evidence.

A few attributes selected for our model go beyond trip level and are on the level of a person. For example, the first trip a person takes during a day. Hence, not every observation is completely independent. To ensure independent observations all people present in the test data will have their trips removed from the training data. By doing this independence is assured. The minimum of five observations rule depends greatly on how the test data is constructed from

the total dataset, which occurs randomly. Hence, we will check each time whether there are enough observations per class with each Chi-square test that will be performed. By doing so both assumptions will be met and the results from the Chi-square tests will be meaningful.

Once a goodness measure has been established the next step is to determine in what situation and using what settings the model performs well. For us it is important that the model is able to predict trip motive for people travelling to a city. The model thus does not only have to perform well on a country level, but also when zoomed in to province and even municipality level. Thus there are three unique situations in which the model has to perform well that differ along the axis of granularity. In terms of settings there are two dimensions. As established in 6.4.2 there are two settings, i.e. adjustable parameters, that will determine how well our final model will perform. These are the weights that are set for m-estimation, i.e. the number of dummy variables taken into account, and the value for minleaf, i.e. the minimum number of observations needed per leaf to continue growing the tree. Figure 6.3 provides an overview of all dimensions along which potential models will be evaluated.

Evaluation will be performed by training the model on a subset of the data and testing the model on the remainder of the data. Chi-square tests will be performed to see how well the model can predict the distribution of trip motives in the test data. To test whether the models perform significantly better than the baseline, i.e. an estimation using the a-priori distribution, the McNemar-test will be performed. The McNemar-test tests if there are differences between two groups based on one dichotomous, i.e. yes or no, dependent variable. Where the dependent paired t-test evaluates continuous variables, the McNemar-test evaluates dependent

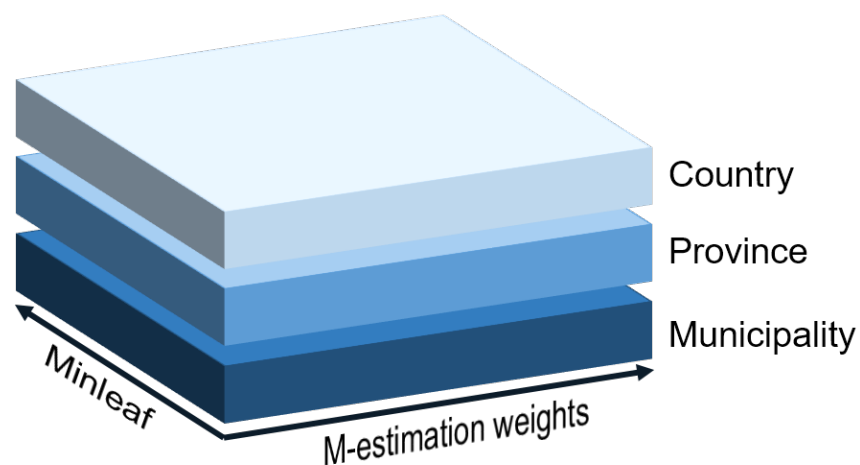


FIGURE 6.3: This figure depicts the axes of evaluation. Potential models will be tested on three levels of granularity and tested using a variety of settings for minleaf and m-estimation weights.

dichotomous variables. In our case the variables are dependent, for example, how well the model predicts trip motives for people going to Amsterdam compared to the baseline model. The outcomes are dichotomous as the distribution is either significantly different or not. McNemar-test thus appears to be a good fit for what we aim to do. [Dietterich \(1996\)](#) also indicated the McNemar-test is one of the most promising statistical tests to test whether one classification algorithm outperforms another. All assumptions for the McNemar-test are also satisfied given our variables are truly dichotomous, i.e. there is no overlap between classes. The McNemar-test will thus provide good insight into how well our model performs and what the added benefit is of employing our model. An alpha level of .05 will be used to test for significance.

6.5.2 Evaluation Results

The results for each level of granularity will be presented in this chapter. In total we evaluated the m-estimation parameter with six unique values and the minleaf parameter with a total of nine values leaving us with a total of 54 combinations. On country level 20 folds are made and evaluated. On this level the a-priori distribution was indistinguishable from the true distribution in each fold. For the different models created we find there is quite a large range in accuracy. The results are shown in table [6.3](#).

TABLE 6.3: Results on country level. Showing how often the model’s predictions are indistinguishable from the true distribution of trip motives. In total 20 Chi-square tests per field are performed.

minleaf	m-estimation					
	0	10	20	30	40	50
0	95%	15%	20%	40%	50%	55%
25	100%	70%	20%	0%	0%	0%
50	100%	100%	65%	55%	20%	0%
75	100%	100%	90%	60%	60%	45%
100	100%	100%	95%	85%	65%	60%
200	100%	100%	100%	100%	95%	90%
300	100%	100%	100%	100%	100%	100%
400	100%	100%	100%	100%	100%	100%
500	100%	100%	100%	100%	100%	100%

In the Netherlands there are a total of 12 provinces. Per combination of minleaf and m-estimation settings we created a model on 11 of the 12 provinces to test how good predictions are on trips going to the other province. The results are shown in table [6.4](#). For reference, the a-priori distribution, which was 100% accurate on country level, is only 50% accurate on

province level. On country level the best models predict 11 out of 12 correct versus 6 out of 12 for the baseline, i.e. a-priori, model. Although we can already see improvements by applying the model they are not significant. The p-value produced by the McNemar-test is .13.

TABLE 6.4: Results on province level. Showing how often the model’s predictions are indistinguishable from the true distribution of trip motives. In total 12 Chi-square tests per field are performed, which is equal to the number of provinces in the Netherlands.

minleaf	m-estimation					
	0	10	20	30	40	50
0	75%	0%	0%	0%	8%	8%
25	75%	25%	8%	8%	8%	8%
50	75%	50%	8%	8%	8%	8%
75	75%	58%	42%	8%	8%	8%
100	75%	75%	50%	33%	17%	8%
200	83%	83%	67%	58%	50%	33%
300	83%	83%	83%	75%	58%	58%
400	83%	83%	83%	83%	75%	67%
500	83%	83%	83%	83%	83%	83%

Although there are many municipalities in the Netherlands, we are unable to test how well the model performs on each one. The assumptions of the Chi-square test are the culprit here. When there are less than 5 observations the test is unreliable. Therefore, we choose to only evaluate the 25 most frequently visited municipalities in our dataset. The a-priori distribution was correct in 19 of the 25 cases (76% accurate). For the evaluation with municipalities the best models do significantly better at predicting the true distribution of trip motives. The McNemar-test produces a p-value of .002, which is greater than the .05 we use for significance. As is shown in table 6.5, the majority of the tested models estimate 22 of the 25 distributions

TABLE 6.5: Results on municipality level. Showing how often the model’s predictions are indistinguishable from the true distribution of trip motives. In total 25 Chi-square tests per field are performed, which is equal to the number of selected municipalities.

minleaf	m-estimation					
	0	10	20	30	40	50
0	76%	72%	72%	76%	76%	76%
25	84%	84%	72%	68%	64%	60%
50	92%	88%	88%	88%	76%	68%
75	88%	88%	88%	88%	88%	88%
100	88%	88%	88%	88%	88%	88%
200	88%	88%	88%	88%	88%	88%
300	88%	88%	88%	88%	88%	88%
400	88%	88%	88%	88%	88%	88%
500	92%	92%	88%	88%	88%	88%

correctly, which is equal to 88%. Two settings excel with 23 out of 25 correct, i.e. minleaf 50 and 500.

6.5.3 Evaluation Conclusion

The key finding here is that on country level the model stays accurate even when targeting more specific users where the a-priori estimations perform significantly worse. Only on a province level both the model and the baseline perform relatively poorly with only 9 out of 12 distributions guessed correctly, i.e. 83%. This is possibly due to the relatively large differences in travel behaviour between the provinces.

Overall we find largely consistent results on all levels in terms of what settings produce good and bad results. On all levels we find that setting minleaf to zero produces relatively bad results. When minleaf is increased results are generally better. However, when the level goes down in terms of area size, a large minleaf starts to produce less accurate results. This might be expected as the model will become more rigid and is less able to focus on subtle differences. M-estimation seems to provide bad results all-round for our dataset. When minleaf becomes larger the negative effects of the m-estimation slightly wear off. However, this is only because the weights have less effect on the probabilities as the number of observations in each leaf are larger. We do not know why the m-estimation produces bad results on country level. Intuitively we would expect that the more weight m-estimation gets the more closely the a-priori distribution is visible. However, we observe the opposite, i.e. the more weight m-estimation gets the worse the model performs.

In terms of finding the best model, in our opinion, there is one clear winner. On each level the model with a minleaf of 500 and no m-estimation produces the best results. The model that will be used for implementation will be the one trained on the entire dataset using these settings. The consequences of the fact that no m-estimation is applied are largely reduced by the fact that a minleaf of 500 is used to construct the model. A minleaf of 500 reduces the need for m-estimation because larger sample sizes have a larger probability of being close to the population distribution, which is exactly what m-estimation is trying to accomplish.

6.5.4 Model Description

The model used for implementation is a [PET](#) trained on the entire dataset with minleaf set at 50. Note that the dataset contains only trips over 10 km in distance, which is the distance as the crow flies. The model is, therefore, only applicable to these trips. It might be used to classify shorter trips, but of those we do not know the accuracy. The design choice was made because trips under 10 km are less well represented in the mobile phone location data (as explained in subsection [5.4.2](#)). Furthermore, these trips are missing not at random, e.g. because they occur within one area, such that a simple weight adjustment would also not be possible. Overall we find the model works exceptionally well. On country level the distributions are almost always estimated correctly in our evaluation. On municipality level accuracies of 92% are found. Only on province level did we find accuracies dropping slightly to 83%. The final model should perform at least as good and probably better as more data is used for training. In addition to more data, there are also no biases in the data, as is the case when leaving out specific provinces and municipalities.

In terms of variable importance we find the arrival time, departure time, and day of the week at the top of the list (see figure [6.4](#)). Runner ups are homebased, i.e. whether someone is going or leaving its home municipality, and trip distance. Together these five attributes make up 85% of the ‘variable importance’. Variable importance is a rather vague measure, nonetheless. Variable importance is measured by observing how many times a variable is used in to make a split in the tree and how often it was a surrogate split, i.e. second or third choice. In figure [6.4](#) the variable importance is scaled to 100%.

Attributes that have been checked but not included in our final model due to disappointing variable importance scores and difficulty with the implementation are: the availability of universities, technical universities at the origin and destination (as a binary value), as well as commercial land use and retail land use at the origin and destination (in acres).

6.6 Implementation

Now that the model has been optimised we will implement the model on the mobile phone location data. Implementing the model is a non-trivial task. We need to make sure the attributes

that the model is trained on match the ones in the mobile phone location data. Furthermore, we need to translate the model from R to an [SQL](#) query.

For nearly all attributes we know the mobile phone location data is similar to the training data. Differences, however, are still possible. In particular, the arrival and departure time of the training data may differ from the mobile phone location data. These are also the most important variables (see figure 6.4) and hence we will have to ensure these match between the both datasets. Departure and arrival times are currently estimated in Mezero by taking the last event at the origin and first event at the destination, respectively.

After implementing the model the distribution of trip motives from [OVIN](#) is compared to Mezero's trip motive distribution. The results are aggregated on a national scale and depicted in figure 6.5. This confirms that the implemented trip motive model is very likely to be implemented correctly. Further results show that also during the morning rush hour the trip motive distributions between the two data sources are very similar (see figure 6.6). We will use the trip motives determined by this model in the results chapter, i.e. chapter 8. However, first the data needs to be prepared for analysis, which is elaborated in the subsequent chapter.

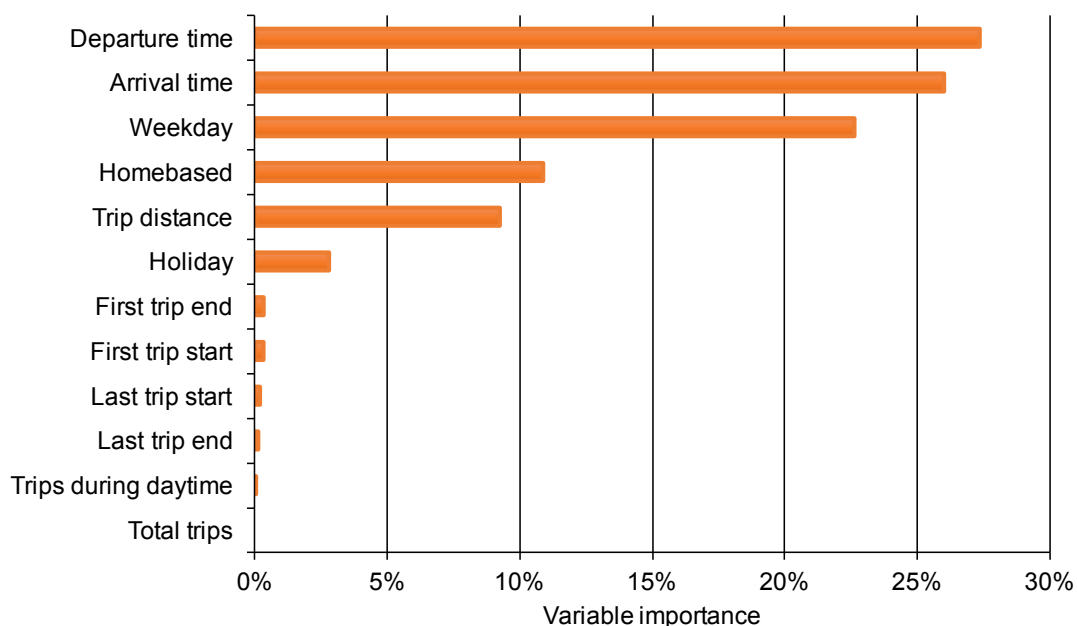


FIGURE 6.4: This figure depicts the relative importance of each attribute that is used in the final model to predict the trip motive. The total variable importance is scaled to 100%.

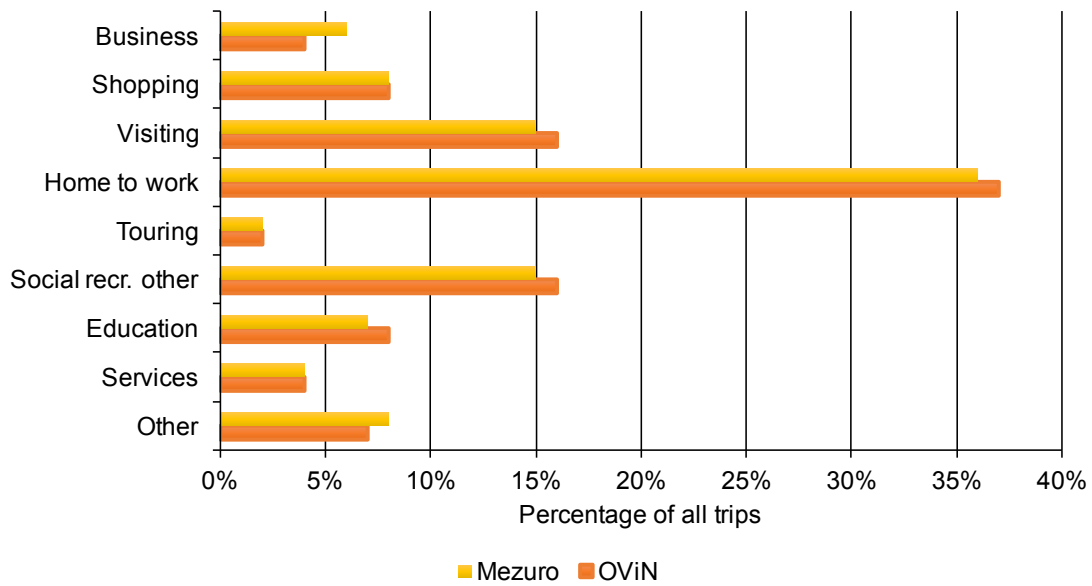


FIGURE 6.5: The comparison of the distribution of trip motives between OViN and Mezero aggregated on the national level.

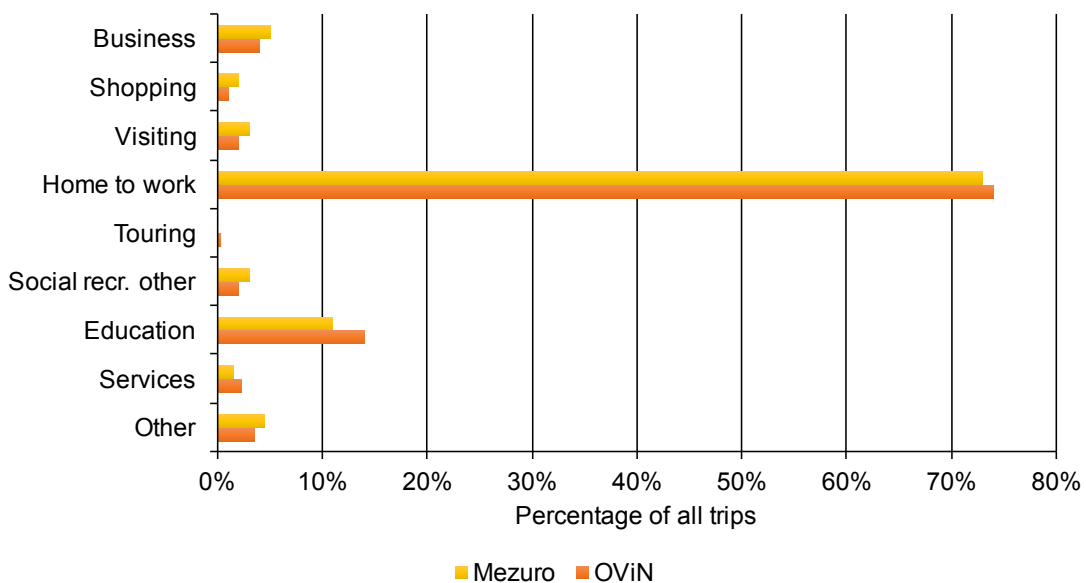


FIGURE 6.6: The comparison of the distribution of trip motives between OViN and Mezero during morning rush hour, i.e. 6am-8am.

Chapter 7

Data Analysis

In this chapter we present how the concept of city attractiveness is operationalised and we select the cities that are included in the analysis. Next to that, we elaborate on the data that is used to represent the factors that influence city attractiveness and how this data has been prepared for analysis. These preliminary steps are necessary to determine the relative importance of factors that influence city attractiveness.

The structure of this chapter is as follows. Section one gives a short introduction. Section two elaborates on the selection procedure for the cities included in the analysis. Section three discusses the factors that are taken into account and data has been collected for. Section four discusses how city attractiveness is measured using mobile phone location data. Section five elaborates on how the data has been prepared for analysis. Finally, section six concludes this chapter.

7.1 Introduction

The goal of this chapter can be divided into four folds: (1) determine which cities are included in the analysis, (2) determine how city attractiveness is calculated using mobile phone location data and (3) collect data on the identified factors and (4) prepare the data for analysis. These activities are depicted in figure 7.1, which forms a part of the main deliverable of the present study. Before data is collected we have to determine which cities are included in our analysis, which is done in the next section.

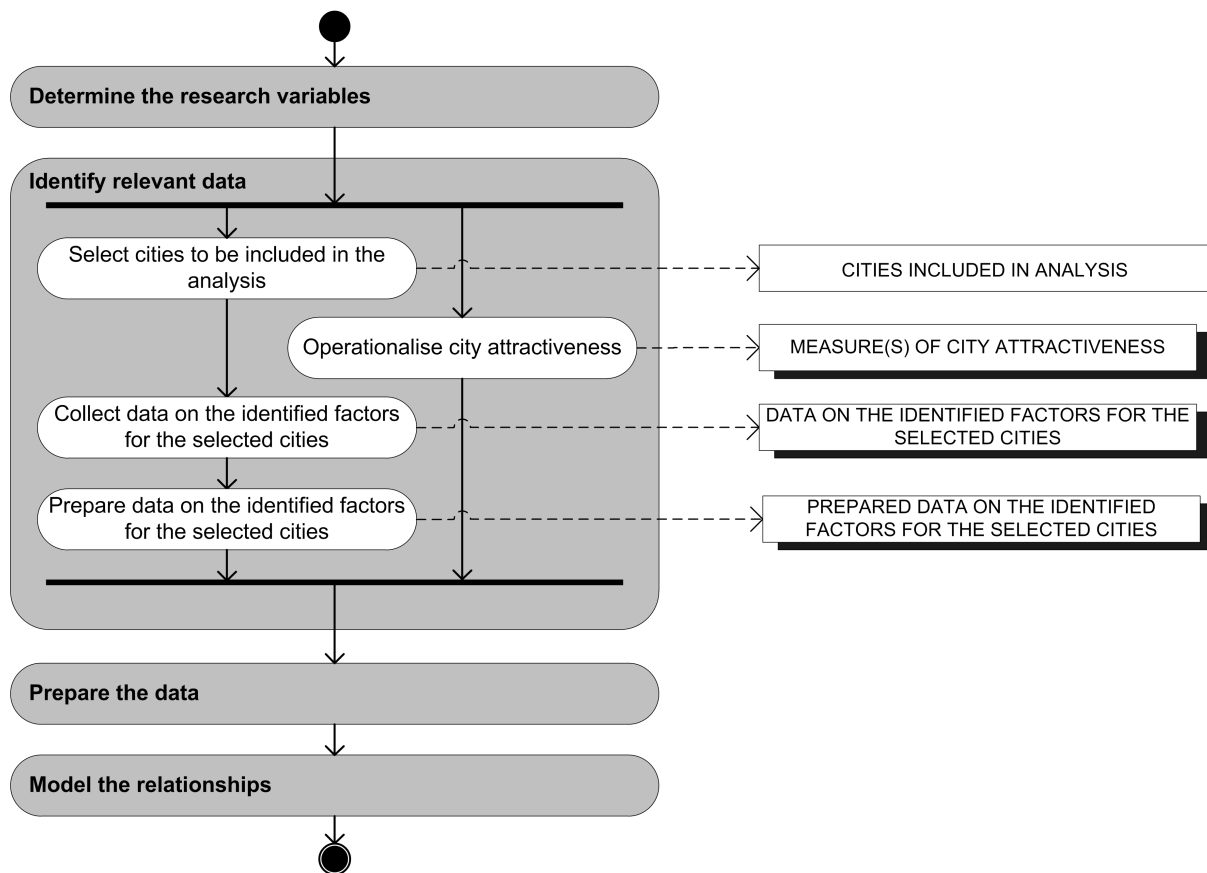


FIGURE 7.1: The second part of the proposed method is applied in this chapter. The other activities in the method are collapsed as these are applied in the chapters 4 and 8.

7.2 City Selection

The selection of cities that are to be included in the analysis is important as these cities are the ones for which data has to be collected on the factors that influence city attractiveness. As briefly explained in our research scope (chapter 2 section 2.2) this study is limited to large cities for two reasons, namely:

1. Visits to large cities can be measured independent of visits to their surrounding areas as opposed to small and medium-sized cities. The Netherlands is subdivided into 1,259 mutually exclusive Mezero areas. Large cities consist of a single Mezero area or multiple Mezero areas, while smaller cities are combined with the urban and rural areas surrounding it. Hence, many small and medium-sized cities cannot be seen independent of their surrounding area making statistics obtained from the mobile phone location data, such as the number of visits, biased.

2. Visits to a large city are less prone to the effects of the minimum-rule-of-15 (one of the privacy limitations discussed in section 5.4.1). This rule implies that OD pairs that contain less than 15 users are omitted and cannot be seen in the data. The impact that this rule has is depicted in figure 7.2. Based on this figure we conclude that less popular Mezuro areas are more likely to be affected by the minimum-rule-of-15 resulting in trips being omitted. Consequently, including small or medium-sized cities, which have less visits than large cities, would lead to a bias at the expense of small and medium-sized cities.

In accordance with our analysis of the impact of the minimum-rule-of-15 we choose the cities to be included in our analysis based on the number of visits they receive to prevent any biases. Figure 7.2 shows that below 100,000 visits the number of trips omitted increases rapidly. Therefore, we use a minimum of 100,000 visits per Mezuro area per month as our cut-off point. If a city consists of multiple Mezuro areas we use the total number of Mezuro areas divided by the total number of visits to those Mezuro areas as our number of visits. In total 30 cities meet this criterion and are therefore included in our analysis (the included cities are listed in table 7.1). How the area that belongs to each city is selected, is elaborated in appendix L.

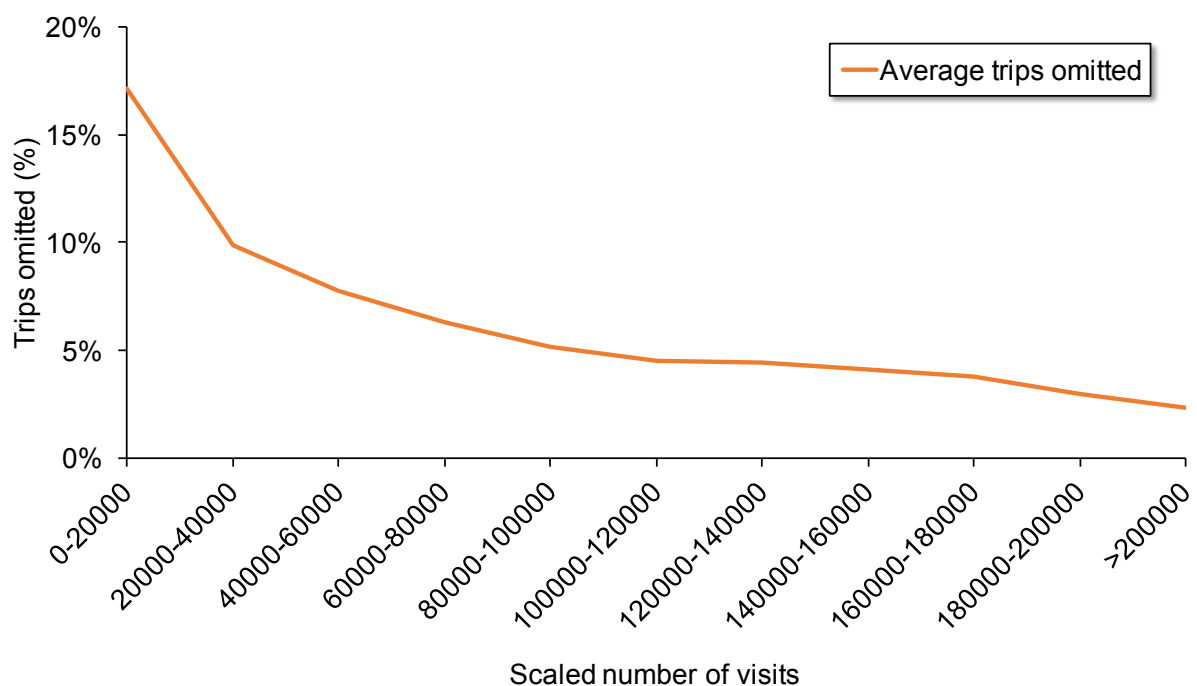


FIGURE 7.2: This figure depicts the impact of the minimum-of-15-rule. The data is based on trips from municipalities to Mezuro areas and is the average per day over the course of a month. The trips omitted is the difference between this level and the province to Mezuro area level. Hence, on average, the more visits a Mezuro area has, the less trips are omitted.

TABLE 7.1: The Dutch cities included in this analysis.

's-Gravenhage	Arnhem	Emmen	Leiden	Tilburg
's-Hertogenbosch	Breda	Enschede	Lelystad	Utrecht
Almere	Delft	Groningen	Maastricht	Venlo
Amersfoort	Deventer	Haarlem	Nijmegen	Zaandam
Amsterdam	Dordrecht	Hilversum	Rotterdam	Zoetermeer
Apeldoorn	Eindhoven	Leeuwarden	Sittard	Zwolle

7.3 Data Collection

In this section we elaborate on the factors that are taken into account in the analysis. The following lists of factors are known to influence city attractiveness and were identified in the literature review presented in chapter 4:

- Determinants that influence the destination attractiveness ([Beerli & Martin, 2004](#))
- Sociodemographic and socioeconomic factors ([Martin et al., 1998](#))
- Determinants of city centre attractiveness ([Teller & Reutterer, 2008](#))
- Competitive forces ([Deas & Giordano, 2001](#))

These lists of factors are elaborated upon in the following four subsections. A complete overview of the descriptive statistics of all included variables that relate to the factors discussed hereafter can be found in appendix J.

7.3.1 Determinants that influence the destination attractiveness ([Beerli & Martin, 2004](#))

The list of determinants by [Beerli and Martin \(2004\)](#) is structured and we use this list to systematically discuss for every factor (1) if it is relevant for our analysis and (2) what data is used to represent the factor. The details of this analysis have been included in appendix I.

To summarise, not every factor in the list by [Beerli and Martin \(2004\)](#) was deemed to be relevant as the list has been proposed as a list of factors that influence destination attractiveness on an international scale. In this analysis, however, we only look at specific cities within the Netherlands. Therefore, some factors have been excluded as the geographical differences in these variables are irrelevant, such as political stability and language barrier. Next to that, the

mobile phone location data is used as a monthly aggregate as daily visitor counts are too much affected by the minimum-rule-of-15. Therefore, atmospheric differences such as the temperature and rainfall, which vary per day, are not taken into account.

Crime rate is included as the average number of reported crimes per 1000 residents of the municipality the city is situated in. Religion is included as (1) the frequency of religious visits per residents per month and (2) the percentage of the population that is religious. Due to a lack of data on city level this is also of the municipality in which the city is situated. The atmosphere of a city, e.g. whether it is luxurious, relaxing or fun, is not included in our analysis as it is hard to quantify using secondary data collection techniques.

All of the factors of the list that were relevant and included are represented using [CBS](#) data. The datasets that are used are: land use, amenities, religion and registered crime rate. We shortly elaborate on amenities to highlight their importance. Amenities, broadly speaking, refer to place-specific assets that are known to contribute to a city's or region's attractiveness ([Öner, 2015](#)). Their importance for city growth and development has been addressed in detail by previous literature ([Ullman, 1954](#); [Clark, Lloyd, Wong, & Jain, 2002](#)). They are a quantification of the characteristics that a city has to offer. For example, the number of cafes within a city, the number of museums in a city and retail property within a city.

7.3.2 Sociodemographic and socioeconomic factors ([Martin et al., 1998](#))

From urban planning literature we know that the variance in the number of trips produced, and therefore city attractiveness, is partly related to socioeconomic and sociodemographic factors of the inhabitants ([Martin et al., 1998](#); [Bruton, 1985](#)). For example, high income households are more likely to produce trips than low income households ([Martin et al., 1998](#)). Hence, cities with a large part part of their inhabitants classified as high income households are more likely to produce trips. Lots of other variables are also known to influence the number of trips generated. To take all these factors into account data was obtained from the [CBS](#).

7.3.3 Determinants of city centre attractiveness ([Teller & Reutterer, 2008](#))

The literature review has shown a possible relation between the accessibility and parking fees on city attractiveness [Teller and Reutterer \(2008\)](#). The first factor is included in our analysis as the distance to the nearest highway. Parking fees are included in the analysis by manually collecting

the average parking fee for every city centre from PrettigParkeren.nl (Prettig Parkeren, 2015). Next to that, the study by Öner (2015) explored the relationship between the access to stores and the place attractiveness of rural and urban environments. They found this relationship to be evident only for urban municipalities and not for rural municipalities, which gives reason to believe that the level of urbanity of a city is related to the attractiveness of a city. Therefore, the level of urbanity is taken into account in our analysis. To express the level of urbanity we use the measure defined by CBS. This measure is based on the address density within a one kilometre radius of each address. The data on this measure is also obtained from CBS.

7.3.4 Competitive forces (Deas & Giordano, 2001)

If we look at the city visitors using the mobile phone location data we see it in hindsight. This means that all people have made a choice in their destination where they go. Some people did not really have a choice in where to go. For example, if someone commutes to his work place he probably has no alternatives in his destination of where to go other than his office. However, there are instances where someone has a choice of where to go. For example, if someone desires to go shopping he can choose the nearest town centre. But perhaps the nearest town centre does not have a wide variety of shops and therefore the person decides to go shopping in a relatively nearby city centre. The fact that other cities perform a certain pull on people has been identified before and is described in the related literature (section 4.3).

To summarise, the supply and attractiveness of alternatives influences the attractiveness of the city. In other words, the more choice potential visitors have, the less attractive their own city becomes. In this research the choice that potential visitors have is named the supply of alternatives. It has been taken into account as depicted in figure 7.3. As we will describe in section 7.5 of this chapter, the census data on the amenities is given for various radii. Depending on the type of amenity it can be a radius of 1 to 5km, 5 to 20km, or 10 to 50km. The supply of alternatives per amenity is quantified as follows. We take the number of amenities of the biggest radius and subtract the number of the amenities within the smallest radius from it. This yields an estimate for the supply of alternatives in the areas surrounding the city. A complete overview of the descriptive statistics of all variables that influence city attractiveness can be found in appendix J.

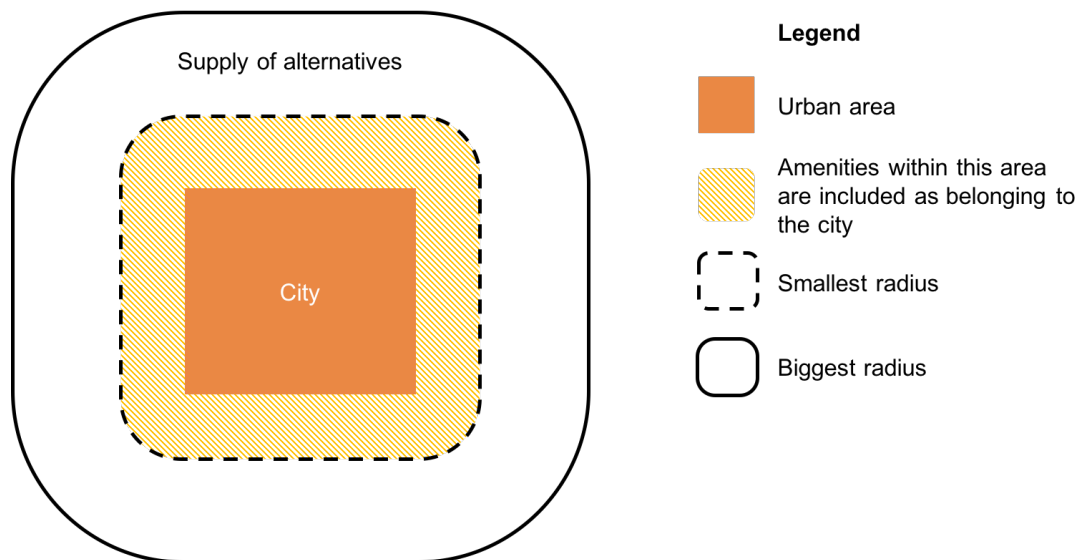


FIGURE 7.3: This figure depicts how the supply of alternatives of the periphery has been determined. The supply of alternatives is calculated by subtracting the number of amenities within the biggest radius from the number of amenities within the smallest radius.

7.4 City Attractiveness

To find out what the relative importance is of factors that influence city attractiveness we must measure city attractiveness. Within the context of this research we defined city attractiveness as *the ability of a city to attract visitors and retain residents*, which is based on the definition of city attractiveness by [Servillo et al. \(2011\)](#) who defined city attractiveness as 'the ability to attract or retain specific target groups'. In this section we develop ways to measure city attractiveness using mobile phone location data.

Arguably, if a city attracts a visitor it offers something that the city of residence of the visitor lacks. Conversely, a resident staying at his city of residence has everything he needs. Based on these two notions we develop a measure that uses a visitor's home location to determine whether trips heading in and out of the city are counted positively, negatively, or not at all towards city attractiveness.

We describe our measure of city attractiveness using the example presented in figure 7.4. The picture shows an example for a person who lives in city depicted in orange. This is important because the city of residence plays an important role in the process of calculating the level of attractiveness per city. For example, when a person lives in Utrecht and travels to Weesp to work there, Weesp offers something to that person that Utrecht didn't offer. Therefore, Weesp's attractiveness increases by 1 and Utrecht's attractiveness decreases by 1. After the person in

our example is done working in Weesp he travels to Amsterdam to visit his friends. This is something Amsterdam offers and therefore this trip increments the attractiveness of Amsterdam. The attractiveness of Weesp is not affected by this outbound trip because Weesp already offered a place to work, which is why the person made the trip in the first place. When the person in question leaves Amsterdam and goes home, his trip does not influence the attractiveness of either city as returning to your city of residence is a logical consequence of travelling out of your city temporarily (this research is scoped to only include temporary migration, see section 2.2). Therefore, trips towards the city of residence are considered not to affect city attractiveness.

We continue by formalising the example given above. All cities are given a unique number j that ranges from 1 to k , where k is equal to the total number of cities. All municipalities are also given a unique number i that ranges from 1 to n , where n is equal to the total number of municipalities. The absolute attractiveness of city j is A_j and can be formally be calculated as the difference between the attracted trips and the non-retained trips of residents, see equation 7.1

$$A_j = I_j - O_j \quad (7.1)$$

where I_j is the total number inbound trips to city j where city j is not the place of residence. This is formally defined in equation 7.2

$$I_j = \sum_{i=1}^n x_{ij} \quad (7.2)$$

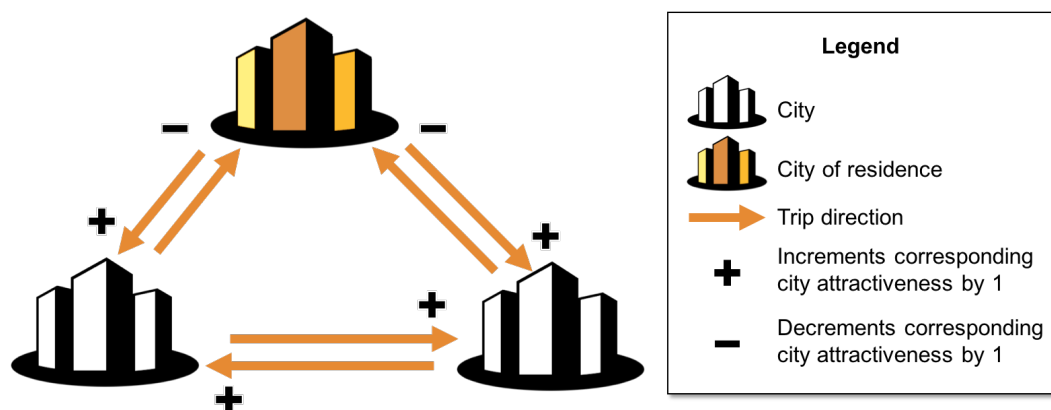


FIGURE 7.4: The effects of the trips performed by a person who lives in the orange-coloured city on the attractiveness for each of the cities involved.

where x_{ij} is the number of trips from municipality i to city j where city j is not the place of residence. F_i is total number of trips departing from city j where city j is the place of residence, as defined in equation 7.3

$$O_j = \sum_{i=1}^n y_{ij} \quad (7.3)$$

where y_{ij} represents the number of trips from city j to municipality i where city j is the place of residence. The absolute measure of city attractiveness presented in equation 7.1 can also be expressed as a value relative to the number of inhabitants of city j (see equation 7.4), in which the number of inhabitants of city j is denoted as S_j . This measure of city attractiveness enables cities to be compared independent of their size.

$$P_j = A_j/S_j \quad (7.4)$$

Another possibility to express the attractiveness of a city, independent of its size, is to use the ratio between the number of attracted and non-retained visits (see equation 7.5). In this case the highest city attractiveness is obtained by maximising I_j and minimising O_j .

$$R_j = I_j/O_j \quad (7.5)$$

The main advantage of using a relative value for the city attractiveness is that the relative importance of factors that influence it can be determined independent of the size of a city. For example, Amsterdam, Utrecht and Rotterdam will always be in the top 3 when calculating A_j , while calculating the R_j and P_j values for these cities will make these cities perform just a little better than average.

7.5 Data Preparation

This section elaborates on the steps that have been taken to prepare the data that has been collected for analysis. In the amenities dataset obtained from the CBS there are two types of measures that present the information per amenity, namely: (1) as the average distance of all addresses within the area to the nearest amenity and (2) as the average number of amenities

within a certain radius of all addresses within the area. For each category (e.g. cinemas or hospitals) of amenities, a measure of the first type is always available. The second type of measure, i.e. the average number of amenities, is preferred but not always available. The reason the second measure is preferred is that it is a better representation of the variety. The distance is a very simplistic measure that hides information compared to the amount of amenities within the area. Therefore, it is expected that the number of amenities is better at explaining the variation in city attractiveness than the average distance to the nearest amenity of a certain category.

Most of the data on cities from the [CBS](#) is on the level of neighbourhoods. Cities consist of multiple neighbourhoods. In order to represent a single value per variable for each city the neighbourhood data needs to be aggregated. Consequently, absolute values, such as the number of men and women, are summed per city. Percentual values and averages, however, are averaged and weighted on the number of inhabitants. In other words, areas with relatively many inhabitants receive additional importance in the calculation of the average, which makes sense. To exemplify, take a city with 2 neighbourhoods: a dockyard with 10 inhabitants and a residential area with 10,000 inhabitants. In the dockyard area the nearest primary care facility is on average 10 kilometres away for all 10 inhabitants. The nearest primary care facility in the residential area, however, is on average 1 kilometre away for each of the 10,000 inhabitants. Without using the weighted mean this would result in an average distance of 5.5 kilometres to the nearest primary care facility. However, this is not a good representation of the reality. Therefore, we use the number of inhabitants of each area as the weight for determining the average of the values, which in case of this example results in a mean of 1.009 kilometres to the nearest primary care facility for the entire city.

Several different data sources are used from different years. [Table 7.2](#) provides an overview of the data sources that have been used and the year the data is from. Our year of preference for our data sources is 2013 because this year is the best trade-off in terms of the availability of data on various topics (such as crime and income levels) and how recent the data is.

7.6 Conclusions

In this chapter the preliminary steps were taken to prepare the data for the modelling phase. We (1) determined which cities are included in the analysis, (2) determined how city attractiveness

TABLE 7.2: These data sources and tables are used in mapping the factors that influence the attractiveness of a destination to data to represent those factors.

ID	Dataset	Source	Year	Table name
1	neighbourhood	CBS	2013	Kerncijfers wijken en buurten
2	land use	CBS	2010	Bodemgebruik; wijk- en buurtcijfers
3	amenities	CBS	2013	Nabijheid voorzieningen; afstand locatie, wijk- en buurtcijfers
4	religion	CBS	2013	Kerkelijke gezindte en kerkbezoek
5	registered criminality	CBS	2013	Geregistreerde criminaliteit; soort misdrijf en regio
6	mobile phone location data	Mezuro	2015	October 2015
7	parking fees	PP.nl	2015	PrettigParkeren.nl

is calculated using mobile phone location data and (3) collected the data for analysis and (4) and prepared the data for analysis. In the next chapter this data will be used to determine the relative importance of the factors that influence city attractiveness.

Chapter 8

Discussion & Results

This chapter presents the results of this research. The goal of this chapter is two-fold. The first goal is to determine the relative importance of factors that influence city attractiveness using mobile phone location data and to discuss these findings. The second goal is to construct a method that can be used to obtain the same results to answer the main research question of this study. The part of this method which is performed in this chapter, is depicted in figure 8.1.

This chapter is structured as follows. Section 8.1 presents the results of a PCA. Section 8.2 discusses the results of this PCA. Section 8.3 presents the attractiveness per city according to the three measures of city attractiveness. Section 8.4 presents the results of regressing the derived PCs on the measures of city attractiveness. Section 8.5 presents the result per trip motive. Section 8.6 discusses the findings of this study. Finally, 8.7 proposes the final deliverable of this study.

8.1 Principal Component Analysis

In the previous chapter a lot of data was prepared on the factors that influence city attractiveness. To determine the relative importance of these factors we use multiple regression analysis. However, multiple regression analysis requires the independent variables in the model to be uncorrelated. Therefore, the first goal is to overcome the problem of multicollinearity and the second goal is to reduce the data complexity.

For this PCA is preferred over factor analysis because factor analysis assumes no multicollinearity while this is not a problem for PCA (Field, 2009). When performing a PCA there are two

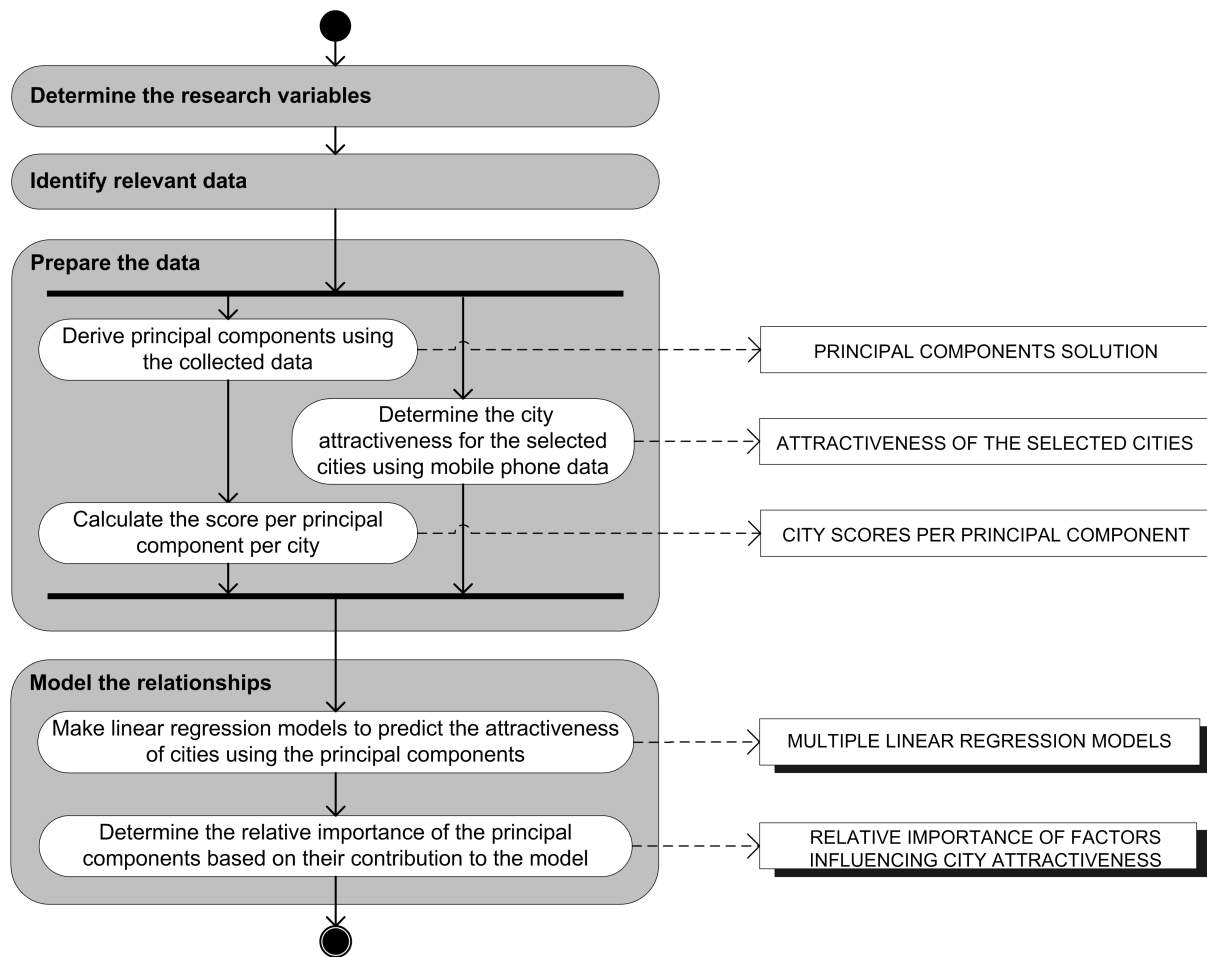


FIGURE 8.1: The final part of the proposed method is applied in this chapter. The other activities in the method are collapsed as these were applied in earlier chapters 4 and 7.

major settings that can be adjusted, i.e. (1) the number of components to retain and (2) the type of rotation to use. First, the number of components to retain is determined using Velicer (1976)'s MAP score, which is chosen for its ability to handle a large number of variables per component (on average >8 variables per component) (Zwick & Velicer, 1986). Figure 8.2 shows that the number of PCs to retain, based on the minimum MAP score, is five. Second, orthogonal rotation is used because the the maximum inter-component correlations when using oblique rotation is .24, which is lower than the threshold of .32 as stated by Tabachnick and Fidell (2001). The varimax rotational method is applied, because the result of this type of rotation is easy to interpret as the variables tend to be associated with one (or a just a few) components (Abdi, 2003; Field, 2009).

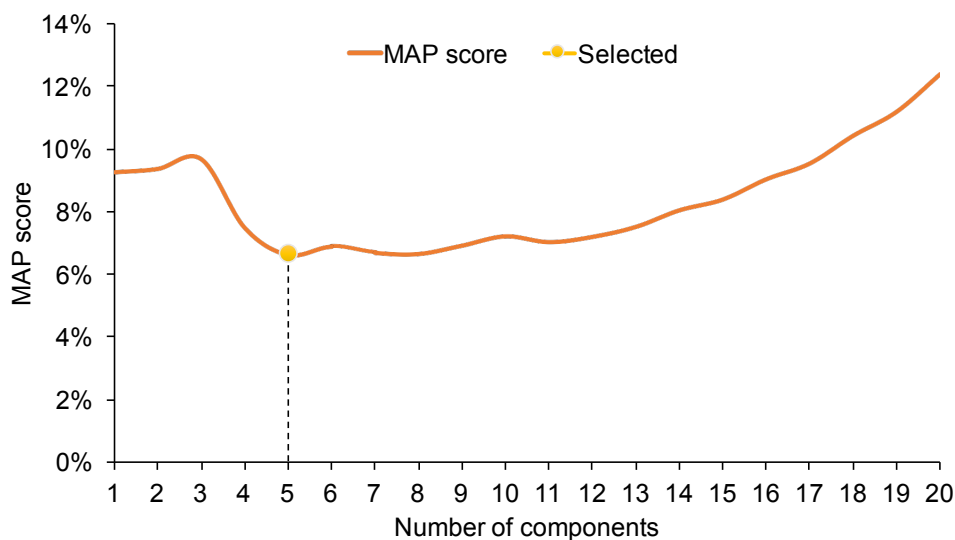


FIGURE 8.2: The number of principal components to retain is determined based on the minimum MAP score. The results indicate that the optimum number of principal components to retain is five.

TABLE 8.1: The results of the PCA. For each variable it is noted how it is loaded onto each of the five PCs. Loadings greater than .6 are in bold. The communality is the amount of shared variance. The uniqueness is a measure of the unique variance explained by the variable. The complexity is the Hofmann's index of complexity, which indicates on average how many components are used to explain each variable (Hofmann, 1978).

Variable	Principal component					Comm	Uniq	Comp
	1	2	3	4	5			
Inhabitants	0.90	0.04	0.06	-0.26	-0.12	0.89	0.11	1.22
Men	-0.10	0.22	-0.30	-0.35	-0.07	0.28	0.72	2.97
Women	0.10	-0.22	0.30	0.35	0.07	0.28	0.72	2.96
Birth rate	0.45	-0.10	0.42	-0.58	-0.08	0.73	0.27	2.90
Death rate	-0.06	0.04	-0.26	0.87	-0.05	0.84	0.16	1.20
Households	0.51	0.08	-0.47	-0.08	0.67	0.94	0.06	2.81
Houses	0.65	0.14	-0.47	0.35	0.22	0.84	0.16	2.84
Income earners	-0.26	-0.13	0.20	0.60	0.14	0.50	0.50	1.87
Unemployment rate	-0.17	-0.50	-0.18	0.10	-0.48	0.54	0.46	2.58
Old-age pension rate	-0.22	0.07	-0.17	0.91	-0.08	0.92	0.08	1.22
Business establishments	0.93	-0.01	0.14	-0.18	-0.12	0.94	0.06	1.15
A Agriculture, forestry and fishing	0.01	-0.16	-0.21	0.03	-0.56	0.38	0.62	1.46
B-F Industry and energy	-0.33	0.44	-0.16	0.08	-0.54	0.64	0.36	2.90
G+I Wholesale and retail trade	-0.41	-0.30	-0.51	0.53	-0.24	0.85	0.15	3.95
H+J Transport, information and communication	0.19	0.01	0.24	-0.66	0.15	0.55	0.45	1.57
K-L Finance and real estate	-0.08	-0.34	0.15	0.66	-0.10	0.59	0.41	1.72
M-N Business services	0.29	0.22	0.57	-0.38	0.49	0.83	0.17	3.68
R-U Culture, recreation and other services	0.57	-0.19	-0.12	-0.28	0.45	0.65	0.35	2.78
Cars	-0.62	-0.30	0.37	0.46	-0.18	0.85	0.15	3.33

Continued on the next page

TABLE 8.1: (continued)

Variable	Principal component					Comm	Uniq	Comp
	1	2	3	4	5			
0-15 years	-0.21	-0.21	0.58	-0.33	-0.58	0.87	0.13	3.12
15-25 years	0.07	0.00	-0.57	-0.38	0.63	0.87	0.13	2.68
25-45 years	0.59	-0.06	0.23	-0.62	0.23	0.84	0.16	2.59
45-65 years	-0.45	0.19	0.12	0.51	-0.53	0.80	0.20	3.33
65 years and older	-0.20	0.07	-0.19	0.92	-0.06	0.92	0.08	1.20
Single	0.54	-0.03	-0.13	-0.60	0.55	0.97	0.03	3.08
Married	-0.63	-0.03	0.24	0.47	-0.52	0.94	0.06	3.19
Divorced	0.19	0.32	-0.16	0.32	-0.67	0.70	0.30	2.27
Widowed	-0.27	-0.01	-0.24	0.88	-0.10	0.91	0.09	1.39
Single households	0.59	0.10	-0.38	-0.13	0.66	0.96	0.04	2.75
Households without kids	-0.63	-0.18	0.07	0.59	-0.35	0.90	0.10	2.77
Households with kids	-0.47	-0.04	0.46	-0.12	-0.69	0.92	0.08	2.67
Average household size	-0.52	-0.12	0.47	-0.02	-0.65	0.94	0.06	2.88
Average house value	0.22	-0.01	0.85	0.02	0.32	0.87	0.13	1.42
Single-family housing	-0.82	-0.41	0.12	0.20	-0.21	0.93	0.07	1.80
Multi-family housing	0.82	0.40	-0.12	-0.19	0.20	0.93	0.07	1.79
Occupied housing	-0.64	-0.11	0.31	-0.34	-0.04	0.63	0.37	2.14
Vacant housing	0.64	0.11	-0.31	0.34	0.04	0.63	0.37	2.14
Private housing	-0.74	-0.31	0.31	0.04	-0.39	0.89	0.11	2.29
Rented housing	0.75	0.34	-0.28	-0.05	0.36	0.88	0.12	2.24
Average electricity usage	-0.73	-0.29	0.17	0.17	-0.42	0.85	0.15	2.24
Average natural gas usage	-0.15	-0.14	-0.07	0.73	0.11	0.59	0.41	1.23
Average income per income earner	0.21	0.22	0.88	-0.16	0.00	0.90	0.10	1.31
Average income per resident	0.18	0.19	0.92	0.04	0.07	0.93	0.07	1.18
Low income residents	0.14	-0.20	-0.90	0.25	0.17	0.96	0.04	1.40
High income residents	0.14	0.25	0.85	-0.28	0.10	0.89	0.11	1.51
Inactive residents	0.21	-0.05	-0.87	-0.07	0.34	0.92	0.08	1.45
Low income households	0.53	-0.04	-0.77	0.12	0.18	0.93	0.07	1.97
High income households	-0.16	0.19	0.89	-0.21	0.08	0.91	0.09	1.29
Urbanity	0.88	0.33	0.07	-0.25	0.09	0.95	0.05	1.50
Crime rate	0.72	-0.10	-0.18	-0.06	0.32	0.67	0.33	1.58
Frequency of religious visits	0.07	-0.09	0.07	-0.13	-0.42	0.21	0.79	1.40
Religious population	0.03	-0.19	-0.07	0.44	-0.01	0.23	0.77	1.45
Parking fees	0.69	-0.15	0.25	-0.14	0.20	0.62	0.38	1.65
Traffic area	0.31	0.28	0.21	-0.18	0.58	0.58	0.42	2.62
Railroad area	0.67	0.22	0.32	0.04	0.26	0.68	0.32	2.04
Road area	0.07	0.29	0.17	-0.26	0.52	0.46	0.54	2.42
Airport area	0.02	-0.16	-0.16	0.10	0.09	0.07	0.93	3.22
Urban area	0.33	0.70	0.10	0.22	0.41	0.83	0.17	2.41
Residential area	0.13	0.73	0.15	0.21	0.40	0.77	0.23	1.93
Retail and hospitality industry area	0.44	0.80	-0.05	-0.02	0.16	0.86	0.14	1.68
Public facilities area	0.32	0.05	0.23	-0.08	0.32	0.26	0.74	2.96
Sociocultural facilities area	-0.04	0.44	-0.05	0.09	0.68	0.68	0.32	1.77

Continued on the next page

TABLE 8.1: (continued)

Variable	Principal component					Comm	Uniq	Comp
	1	2	3	4	5			
Commercial area	0.57	0.26	-0.03	0.20	0.10	0.45	0.55	1.78
Recreational area	0.02	0.85	-0.07	-0.21	0.20	0.81	0.19	1.25
Parks and greenspace area	-0.03	0.80	-0.06	-0.22	0.14	0.71	0.29	1.23
Sports area	0.04	0.64	-0.01	0.03	0.43	0.60	0.40	1.76
Rural area	-0.49	-0.66	-0.23	0.01	-0.09	0.74	0.26	2.16
Countryside area	-0.49	-0.66	-0.23	0.00	-0.08	0.73	0.27	2.16
Forest and other natural area	-0.22	-0.37	0.15	0.02	-0.38	0.35	0.65	2.91
Forest area	-0.25	-0.41	0.16	0.10	-0.24	0.32	0.68	2.94
Dry natural area	0.12	-0.10	0.16	0.04	-0.29	0.14	0.86	2.27
Wet natural area	-0.22	-0.13	-0.07	-0.32	-0.57	0.50	0.50	2.06
Inland water area	0.50	0.13	0.00	-0.26	-0.31	0.43	0.57	2.43
Primary care facilities	0.88	0.28	-0.01	-0.14	0.06	0.87	0.13	1.27
Hospitals	0.75	0.18	-0.12	-0.21	0.06	0.65	0.35	1.36
Supermarkets	0.90	0.20	0.03	-0.22	0.16	0.92	0.08	1.29
Other retail	0.92	0.25	0.04	-0.10	0.04	0.93	0.07	1.18
Department stores	0.89	0.10	0.13	-0.02	0.12	0.84	0.16	1.11
Cafes	0.92	0.19	-0.08	-0.06	0.18	0.93	0.07	1.19
Cafeterias	0.89	0.22	0.03	-0.10	0.30	0.93	0.07	1.38
Restaurants	0.86	0.18	0.10	-0.12	0.27	0.87	0.13	1.36
Hotels	0.86	-0.05	0.10	-0.02	-0.04	0.75	0.25	1.04
Highway	0.20	-0.08	-0.09	0.25	-0.26	0.19	0.81	3.35
Railway station	0.14	-0.19	-0.04	0.07	-0.22	0.11	0.89	3.05
Libraries	-0.42	-0.17	0.00	0.03	-0.45	0.41	0.59	2.29
Pools	-0.29	-0.32	-0.17	0.42	-0.16	0.42	0.58	3.44
Ice rinks	-0.17	-0.29	0.00	0.21	-0.35	0.28	0.72	3.16
Museums	0.94	0.18	0.10	-0.12	0.08	0.94	0.06	1.14
Cinemas	0.77	0.01	0.18	-0.15	0.39	0.80	0.20	1.70
Amusement parks	0.62	0.19	-0.13	-0.07	-0.05	0.45	0.55	1.32
Podium arts	0.94	-0.05	0.14	-0.10	-0.01	0.91	0.09	1.07
Peripheral amusement parks	0.26	0.53	0.52	-0.32	0.00	0.73	0.27	3.13
Peripheral cinemas	0.25	0.78	0.31	-0.11	0.03	0.78	0.22	1.58
Peripheral cafes	0.98	0.04	0.00	-0.05	-0.04	0.97	0.03	1.01
Peripheral cafeteria	0.97	0.09	0.08	-0.11	0.00	0.96	0.04	1.06
Peripheral hotels	0.24	0.74	0.06	0.11	-0.02	0.62	0.38	1.27
Peripheral primary care facilities	0.91	0.25	0.04	-0.15	-0.05	0.91	0.09	1.22
Peripheral museums	0.25	0.83	0.16	-0.17	0.03	0.81	0.19	1.36
Peripheral podium arts	0.25	0.79	0.24	-0.15	-0.03	0.77	0.23	1.49
Peripheral restaurants	0.96	0.04	0.12	-0.11	-0.01	0.95	0.05	1.06
Peripheral supermarkets	0.93	0.15	0.07	-0.19	0.01	0.93	0.07	1.15
Peripheral department stores	0.28	0.85	0.20	-0.19	0.06	0.88	0.12	1.46
Peripheral other retail	0.95	0.17	0.08	-0.11	-0.09	0.96	0.04	1.12
SS loadings	30.59	11.78	11.14	10.67	10.17			

The five resulting **PCs** explain 72% of the variance present in the original dataset. Based on the sum of the squared loadings it can be concluded that each **PC** accounts for 41%, 16%, 15%, 14% and 14% of the variance explained, respectively. This implies that the order of importance of the **PCs** decreases from left to right.

The advantage of performing a **PCA** is that we now have five **PCs**, i.e. composite variables, that show very little mutual correlation ($R < .01$). Because one of the assumptions of multiple linear regression is that there is no correlation between the independent variables, the **PCs** derived using **PCA** make it possible to perform a multiple linear regression analysis using 72% of the variance of the original dataset. Moreover, the reduced amount of variables make it easier to interpret the results. Before we continue to build models using these five **PCs** in order to determine their relative importance, we will interpret and assign names to the **PCs**.

8.2 Principal Components

In this section the five **PCs** identified using the **PCA** are assigned a name and discussed. The names are assigned based on the relationship of the majority of significant variable loadings comprised in the respective **PCs**. The discussion is performed based on the scores of the 30 cities on each of the five **PCs**. A graphical representation of the scores is depicted in figure 8.3. In this figure each combination between a city and a **PC** score is presented as colour. A yellowish colour indicates a high score while a dark blue colour indicates a low score. The factor scores are obtained using Thurstone's regression method (Thurstone, 1935). Tabachnick and Fidell (2001) conclude that the regression method is preferred because it is most easily understood. A score of 0 implies that the city score is the mean, while, for example, a score of -1 implies that the score is one standard deviation below the mean. We will now elaborate on each of the five **PCs** that were found.

8.2.1 City size

When looking at the variable loadings per **PC** in table 8.1 it is noted that the first **PC** has many variables that are related to the size of the city. This is mainly confirmed by the significant loadings on the number of inhabitants (.90) and number of business establishments (.93), urbanity (.88), and many of the amenities within and in the periphery of the city. The first **PC** is

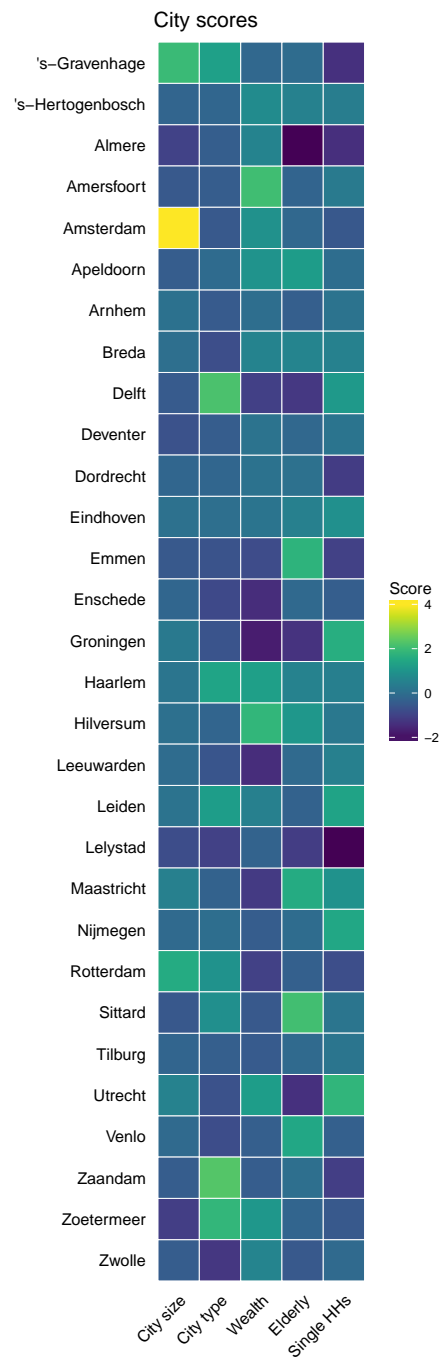


FIGURE 8.3: An overview of the scores per city per PC.

named 'city size' because all the variables with significant loadings on the first PC are known to scale proportionally to the size of a city.

Referring to the scores on the city type figure 8.3, the high scores of major cities such as Amsterdam, Rotterdam and Utrecht confirm that the PC city size is largely dependent upon the size of a city. However, even though the cities Almere, Zoetermeer and Lelystad are number

7, 16 and 25 in terms of the number of inhabitants respectively, they have the lowest scores on the first component, i.e. city size. This can be explained by the absence of many of the amenities, which also have high loadings on the first component (see the variable loadings table 8.1). These cities are the bottom three when looking at the number of cafes per inhabitant and the number of restaurants per inhabitant. Therefore, the component city size can be seen as a measure of the availability of amenities that are related to city size rather than the number of inhabitants only. For example, Almere is pretty large in terms of the number of inhabitants, but has a relatively small ratio of bars, restaurants and museums per inhabitant. Therefore, it scores low on the component city size.

8.2.2 City type

The second PC is loaded with variables that are related to the amount of urban area (.70), residential area (.80), retail and hospitality industry area (.80), recreational area (.85), parks and greenspace area (.80), sports area (.64) and number of amenities within the periphery of the city (.74 to .85). This PC is negatively loaded with variables that are associated with the amount of rural area (-.66) and countryside area (-.66), meaning that the second PC is related to the relative amount of leisure and residential area within the city and the number of amenities available in the periphery of the city. This combination gives an indication of the type of city or perhaps more fitting: its geographic location relative to other cities. The geographic location and the supply of alternatives in the periphery of the city give reason to believe that this PC relates to the 'city type'.

Cities with high scores on the city type component are Delft, Zaandam, Zoetermeer, Leiden and Haarlem. The communality between these cities is that they are all located within the Randstad and near to larger cities such as Amsterdam, Rotterdam and 's-Gravenhage. These large cities have a lot of museums, department stores and hotels. Consequently, cities that are located near to these major cities have relatively large scores on the variables that quantify the availability of amenities within the periphery of the city. For example, Delft, Zoetermeer and Zaandam are the top 3 cities with the most museums in their periphery. Moreover, they score low on the negatively loaded variables rural area and countryside area, which is good in terms of score improvement on this component.

On the other hand, low scores on this component are obtained by Zwolle and Lelystad. Lelystad has the lowest score on all of the significant land use and peripheral amenity variables in this

component. In terms of the negatively correlated variables (rural and countryside area) its scores are in the top 5, meaning the score will be even lower. Zwolle has the highest percentage of rural area and countryside area, which is mainly why it scores the low on this component. Next to that it is located relatively far away from big cities, which have a high density of amenities.

To conclude, cities within the proximity of large cities tend to score high on this component. This is due to two main things: (1) the periphery of these cities contains a high concentration of amenities and (2) a large percentage of the area within these cities is reserved for the purpose of leisure activities such as sports area, parks and greenspace. Based on the city scores and the variable loadings of this PC two types of cities can be distinguished. The highest scores (>2) on this PC indicate that a city functions as a residential city near one or multiple larger cities. Examples include Zoetermeer and Zaandam. Lower scores are obtained by cities that are geographically more isolated by rural area and do not have much residential area, such as Zwolle and Lelystad. The other cities score somewhere around the mean on this PC.

8.2.3 Wealth

The third PC has significant positive loadings on the variables average house value (.85), average income per resident (.92), high income residents (.85) and high income households (.89). The same PC has significant negative loadings on the low income residents (-.90), inactive residents (-.87) and low income households (-.77). Consequently, this PC is related to incomes in a city and will be referred to as 'wealth'.

The significant loadings in this component are all related to the wealth standard of the inhabitants of a city. All variables obtain loadings of .77 to .92, which indicates that these variables are highly related. The cities that score high on this component are Amersfoort, Hilversum and Zoetermeer. These cities have on average 24% high income household, while at the other end of the spectrum, on average only 12% of the households are considered to have high incomes.

8.2.4 Elderly

The fourth PC has significant loadings on the death rate (.87), old-age pension rate (.91), the part of the population aged 65 years and older (.92), the percentage widowed (.88), and average natural gas usage (.73). Therefore, this component is derived from variables that are related to the presence of elderly people within a city and is why this PC is named 'elderly'.

Interestingly, the fifth highest loaded variable, average natural gas usage, is also related to the elderly population as elderly households use significantly more natural gas than the average household (Brounen, Kok, & Quigley, 2012). The highest scores on this PC are obtained by Sittard, Maastricht and Emmen. These cities all have the highest share of inhabitants aged 65 years and older in their population, which is around 20%. Almere, scoring the lowest on this component, only has 8% of its inhabitants aged 65 years or older.

8.2.5 Single households

The fifth and final PC has significant loadings on the number of households (.67), single households (.66), and the population aged 15-25 years (.63). Negative loadings are on the households with kids (-.69) and average household size (-.65). Therefore, a positive score on this component would mean that there are relatively many single households in this city. Conversely a negative score would indicate relatively little single person households. Therefore, this PC is named 'single Households (HHs)'.

The variable loadings on this component are not highly loaded, meaning the significant variables are loosely coupled. A pattern can be identified because the positively loaded variables are related to the number of households per inhabitant, which is related to the percentage of single households. On the contrary, the significant negative loadings are related to their counterparts, i.e. households with kids and household size.

The cities in Flevoland, which are included in this analysis, i.e. Almere and Flevoland, score very low on this PC. They have large household sizes and little sociocultural facilities area. Sociocultural facilities area includes, but is not limited to, area occupied by facilities such as churches, theatres, cinemas and hospitals. The percentage of the population aged 15-25 is relatively high in these cities. In figure 8.3 the score on this component is inversely proportional to the household size. So, if the average household size is large, the score will be blueish. If the average household size is small and there are lots of single households, the score is high.

Now that we have uncorrelated PCs that can be used as independent variables in our multiple linear regression analysis, we will determine values for our dependent variable, i.e. city attractiveness, in the next section.

8.3 Determining City Attractiveness

The three measures of city attractiveness that were proposed in chapter 7 section 7.4 are determined for each city using the mobile phone location data. The resulting scores are presented in table 8.2. This table shows that the absolute measure of city attractiveness (A_j) is very similar to the population size of each city. A correlation of .90 between the number of inhabitants and city attractiveness (A_j) confirms that a large population is highly related to more in- and outbound traffic.

The second measure of city attractiveness (P_j) shows a very different ranking of cities. This measure shows the city attractiveness (A_j) relative to the number of inhabitants of a city. When correcting the absolute city attractiveness for the number of inhabitants it can be seen that some regional cities perform the best. This can be explained by the fact that these cities do not have a lot of outbound traffic of their inhabitants, because the nearest city is very far away. Consequently, these cities are good at retaining their inhabitants. Additionally, these

TABLE 8.2: The attractiveness of each city per measure of city attractiveness.

Rank	City name	Attractiveness (A_j)	City name	Attractiveness per inhabitant (P_j)	City name	Attracted per non-retained (R_j)
1	Amsterdam	9421693	Venlo	4.0	's-Hertogenbosch	16.2
2	Rotterdam	5324198	Eindhoven	3.7	Venlo	15.6
3	Utrecht	4683619	Groningen	3.5	Breda	15.5
4	's-Gravenhage	3294745	Leeuwarden	3.4	Eindhoven	14.8
5	Eindhoven	3240358	Zwolle	3.3	Zwolle	14.7
6	Groningen	2401931	Breda	3.3	Utrecht	14.5
7	Breda	2277823	Utrecht	3.3	Amsterdam	13.2
8	Arnhem	1962865	Amsterdam	3.2	Arnhem	13.1
9	's-Hertogenbosch	1843780	's-Hertogenbosch	3.1	Groningen	12.3
10	Zwolle	1806575	Maastricht	3.0	Leeuwarden	11.7
11	Nijmegen	1317889	Arnhem	2.9	Rotterdam	9.5
12	Amersfoort	1202659	Sittard	2.7	Sittard	9.4
13	Tilburg	1092154	Rotterdam	2.7	Maastricht	8.7
14	Leeuwarden	1068290	Emmen	2.5	Delft	8.3
15	Maastricht	1055334	Nijmegen	2.4	Lelystad	8.2
16	Venlo	1045247	Delft	2.3	Amersfoort	8.0
17	Apeldoorn	837216	's-Gravenhage	2.3	Nijmegen	7.9
18	Delft	822131	Amersfoort	2.0	Emmen	7.8
19	Leiden	818866	Enschede	2.0	Leiden	6.8
20	Enschede	734866	Apeldoorn	2.0	's-Gravenhage	6.5
21	Lelystad	621647	Leiden	2.0	Apeldoorn	6.3
22	Dordrecht	538111	Tilburg	1.9	Hilversum	6.0
23	Hilversum	514065	Lelystad	1.8	Tilburg	5.8
24	Haarlem	479720	Hilversum	1.7	Enschede	4.6
25	Emmen	441521	Dordrecht	1.6	Dordrecht	4.5
26	Sittard	351501	Deventer	1.5	Deventer	3.9
27	Deventer	317992	Haarlem	1.4	Zaandam	3.3
28	Almere	302880	Zaandam	1.4	Haarlem	3.1
29	Zaandam	249127	Zoetermeer	1.2	Zoetermeer	2.0
30	Zoetermeer	245763	Almere	1.2	Almere	1.6

cities function as regional centres attracting many of the inhabitants of the surrounding smaller cities and towns. The cities that are at the bottom are considered to be residential cities that provide workers to a larger city nearby. Thereby, the inhabitants of these cities create a lot of outbound traffic, which makes these cities bad at retaining their residents thus obtaining low scores on the city attractiveness per inhabitant (P_j).

The third measure of city attractiveness (R_j) shows the ratio between the attracted visitors and non-retained residents. This ranking shows which cities are the best at attracting visitors relative to how bad they are at retaining their own inhabitants. This ranking is fairly related to the second measure of city attractiveness (P_j) as the correlation between these two rankings is .93.

In the next section we will apply multiple linear regression using these measures of city attractiveness as dependent variables.

8.4 Relative Importance of Factors Influencing City Attractiveness

To determine the relative importance of each PC on city attractiveness multiple linear regression is used. Multiple linear regression is preferred over a simple correlation analysis, because depending on the type of rotation used in the PCA, correlation can exist between the resulting PCs. When no correlation is present the relative contribution of each predictor to the overall model performance is given by the predictor's R^2 . However, when regressors are correlated, it is no longer straightforward to break down model R^2 into shares from the individual regressors and more sophisticated methods should be used Grömping et al. (2006).

In our case the predictor variables used in the multiple linear regression analysis are derived using a PCA with orthogonal rotation. Orthogonal rotation implies that the resulting PCs are uncorrelated (Field, 2009), which means each predictor's contribution can be determined in terms of its R^2 -value (Grömping et al., 2006).

How each dependent and independent variable is related is depicted in the scatterplot matrix in figure 8.4. For each of the dependent variables, i.e. measures of city attractiveness, a multiple linear regression model is constructed. All three models use the PCs as independent variables.

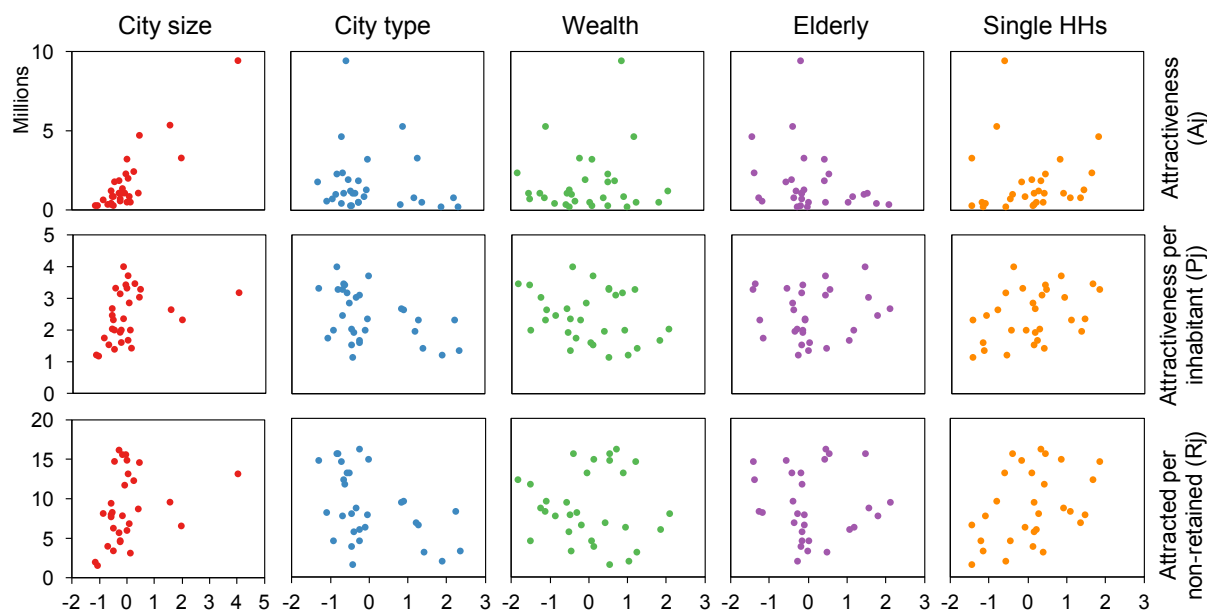


FIGURE 8.4: Scatterplots between the five PCs and the measures of city attractiveness.

To construct the models the `lm()` function that is available in R is used (Chambers, 1991). The results can be found in table 8.3.

For each possible dependent and independent variable combination two values are denoted in table 8.3. The first is the coefficient and the second, which is between parentheses, is the

TABLE 8.3: Regression Results

	<i>Dependent variable:</i>		
	Attractiveness (1)	Attractiveness per inhabitant (2)	Attracted per non-retained (3)
City size	1,731,884.000*** (148,959.500)	1.317* (0.672)	0.282** (0.106)
City type	-276,616.700* (148,959.500)	-2.006*** (0.672)	-0.343*** (0.106)
Wealth	166,185.900 (148,959.500)	-0.244 (0.672)	-0.219* (0.106)
Elderly	-351,894.900** (148,959.500)	0.465 (0.672)	0.153 (0.106)
Single HHs	140,567.300 (148,959.500)	1.620** (0.672)	0.331*** (0.106)
Intercept	1,677,153.000*** (146,455.800)	8.800*** (0.661)	2.441*** (0.105)
Observations	30	30	30
R ²	0.859	0.444	0.583
Adjusted R ²	0.830	0.328	0.496
F Statistic (df=5; 24)	29.268***	3.835**	6.700***

Note:

*p<0.1; **p<0.05; ***p<0.01

standard error. The standard errors of the PCs are the same for each measure because the variables were standardised before performing the PCA.

As expected, the first model performs the best explaining 85,9% of the variance in the dependent variable ($R^2 = .859$, $p < .01$). This is explained because of the strong and significant relationship between the PC city size and absolute city attractiveness (A_j), which contributes for 92.4% to the model's performance. The relative importance of the variables is calculated based on their contribution to the overall R^2 of the model and is denoted in table 8.4.

The second model, where P_j is the dependent variable, has the largest significant contribution by city type (46.5%). What is more, the relationship is inverted, which implies that the higher a city scores on city type, the lower the city performs on this measure of city attractiveness. Recall that residential cities such as Almere, Zoetermeer en Zaandam scored high on this PC. That makes sense because the three aforementioned cities are the lowest three cities on this measure of city attractiveness (see figure 8.3).

The third model also has the highest significant contribution to the overall model performance by the PC city type (31.1%). However, the PC single households is a close second with a contribution of 29.0% to the overall model R^2 .

To make sure that the results are valid the assumptions that apply to linear regression are tested for each of the models. We test for (1) normality of the residuals, (2) homoscedasticity of the error variance, (3) independence of the residuals. The Shapiro-Wilkinson normality test is performed to test for the normality of the residuals and is insignificant for all models ($W = 0.97894$, $p = 0.7968$ for the first model), meaning the assumption of normally distributed residuals is satisfied. Additionally, we test for homoscedasticity of the error variance using the Breusch-Pagan test (Breusch & Pagan, 1979). The test remains insignificant for all models ($\chi^2 = 1.5$, $df = 1$, $p = .22$ for the first model), which indicates that the variances are equal and the assumption of homoscedastic error variance is satisfied. The Durbin-Watson test is performed

TABLE 8.4: The relative importance of the PCs.

	Attractiveness (1)	Attractiveness per inhabitant (2)	Attracted per non-retained (3)
City size	92.4%	20.0%	21.0%
City type	2.4%	46.5%	31.1%
Wealth	0.9%	0.7%	12.7%
Elderly	3.8%	2.5%	6.2%
Single HHs	0.6%	30.3%	29.0%

to test for the presence of autocorrelation in the residuals, i.e. if the residuals are independent from one another. The Durbin-Watson tests are insignificant for all models ($D-W = 1.53$, $p = .23$ for the first model), which implies that the residuals are independently distributed and that this assumption is satisfied. Therefore, we can conclude that the results produced by the linear regression models are valid.

8.5 Relative Importance per Trip Motive

The three measures of attractiveness that were used to determine the relative importance of the five PCs can also be determined per trip motive using the trip motive algorithm that was developed and implemented within the context of this research (see chapter 6). For each trip motive (9 in total) and each measure of attractiveness (3 in total) a multiple linear regression model has been fitted. This means 27 linear regression models are created with the five principal components included as predictor variables.

The difference between the previous models from section 8.4, is that in this case the dependent variable is calculated per trip motive. The three measures of city attractiveness remain the

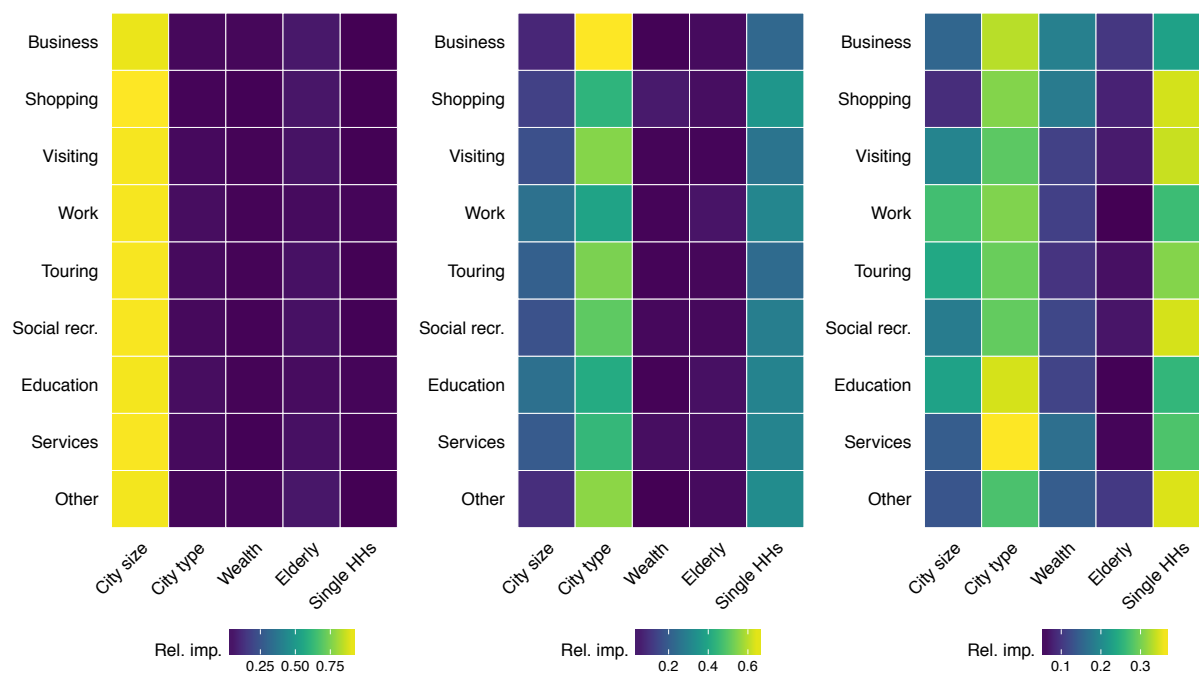


FIGURE 8.5: Relative importance of the five PCs per trip motive per measure of attractiveness.

same, but using the trip motive model that was implemented within the context of this research it is possible to determine a city's attractiveness for a specific trip motive.

The relative importance per trip motive is depicted in figure 8.5, where the left figures from left to right use A_j , P_j and R_j as the dependent variable. The tables containing the quantitative results can be found in appendix K. The results of the linear models used to obtain the relative importance can be found in appendix H.

What can be noted from the results in figure 8.5 is that the first measure of city attractiveness shows barely any deviation from the mean per trip motive. The second measure shows that trips with a business motive are more affected by city type than city size. The third measure shows that shopping trips are less affected by city size. We have not been able to find a solution as to why we observe these differences in the relative importance trip motive. This could be due to the fact that the trip motive algorithm does not accurately capture the differences per city as cities deviate from the mean (see the limitations of this research in section 9.3).

8.6 Discussion

As mentioned in the literature review on urban competitiveness (section 4.2), performed in the first part of the proposed method, it was identified that city attractiveness is influenced by competitive forces (Sinkienė, 2008). Recall that this was the primary reason for constructing and including the peripheral variables in the analysis (section 7.3.4). After conducting the PCA the variable loadings (table 8.1) show that the resulting component city type includes many significant variable loadings on these peripheral variables, which implies it is strongly related to the the competitive forces.

Looking at the table 8.1 not all of the peripheral variables are included within the PC city type. Some are also included in the PC city size. This is explained by the data that was obtained. As described in 7.3.4 the density of amenities depends on the type of amenity and can be 1km to 5km, 5km to 20km, or 10km to 50km. The peripheral variables of the smallest category (1km to 5km) are included in the city size, while the variables of the 5km to 20km are included in the second PC city type. Peripheral attractions, the only variable in the 10km to 50km category, was insignificant (for more information see the descriptive statistics of the peripheral variables in Appendix J). Consequently, this strengthens our belief that the second PC describes the geographically location of the city relative to other cities.

Apart from the absolute measure of city attractiveness (A_j), city type is the most important component that influences city attractiveness. In both the city attractiveness per inhabitant measure (P_j) and the ratio measure of city attractiveness (R_j) the component city type explained 46% and 31% of the variance in the attractiveness of a city, respectively (see table 8.4). This also shows that the relative importance of factors that influence city attractiveness is highly dependent upon how city attractiveness is defined.

Comparing the results of our application of the proposed method to existing literature in order to evaluate the results is not possible. The problem is that no single study has been found that applies a similar scope to the one that was used in this research. For example, in the tourism research only the relative importance of the factors that influence a tourist's destination image are studied (Beerli & Martin, 2004). The factors they evaluated can be considered merely a subset of the factors that were evaluated in this study. Moreover, in the city centre attractiveness literature the same has been researched but this only applies to factors that influence city centre attractiveness (Teller & Reutterer, 2008). As most of these studies are performed using surveys they did not include a wide variety of factors to limit the time taken to fill in these surveys. Because the PCs identified within this research are considered to be very general and our measure of city attractiveness as well it is hard to compare the results. However, the results we in the general models found in section 8.4 are considered to be logical. The trip motives do show results, but the differences we are not able to validate these differences by using existing research because the PCs identified are too generic and have not been used by existing research.

8.7 Proposed Method

This research set out to assess how the relative importance of factors that influence city attractiveness can be determined using mobile phone location data. At the start of chapter 4, 7 and this chapter, method fragments have been depicted and applied. These method fragments are part of the full method that forms the main deliverable of this study (see figure 8.6). This method can be applied to obtain the relative importance of factors that influence city attractiveness using mobile phone location data as shown in this research. The activity table and concept table that elaborate the activities and concepts used in the proposed method can be found in Appendix B.

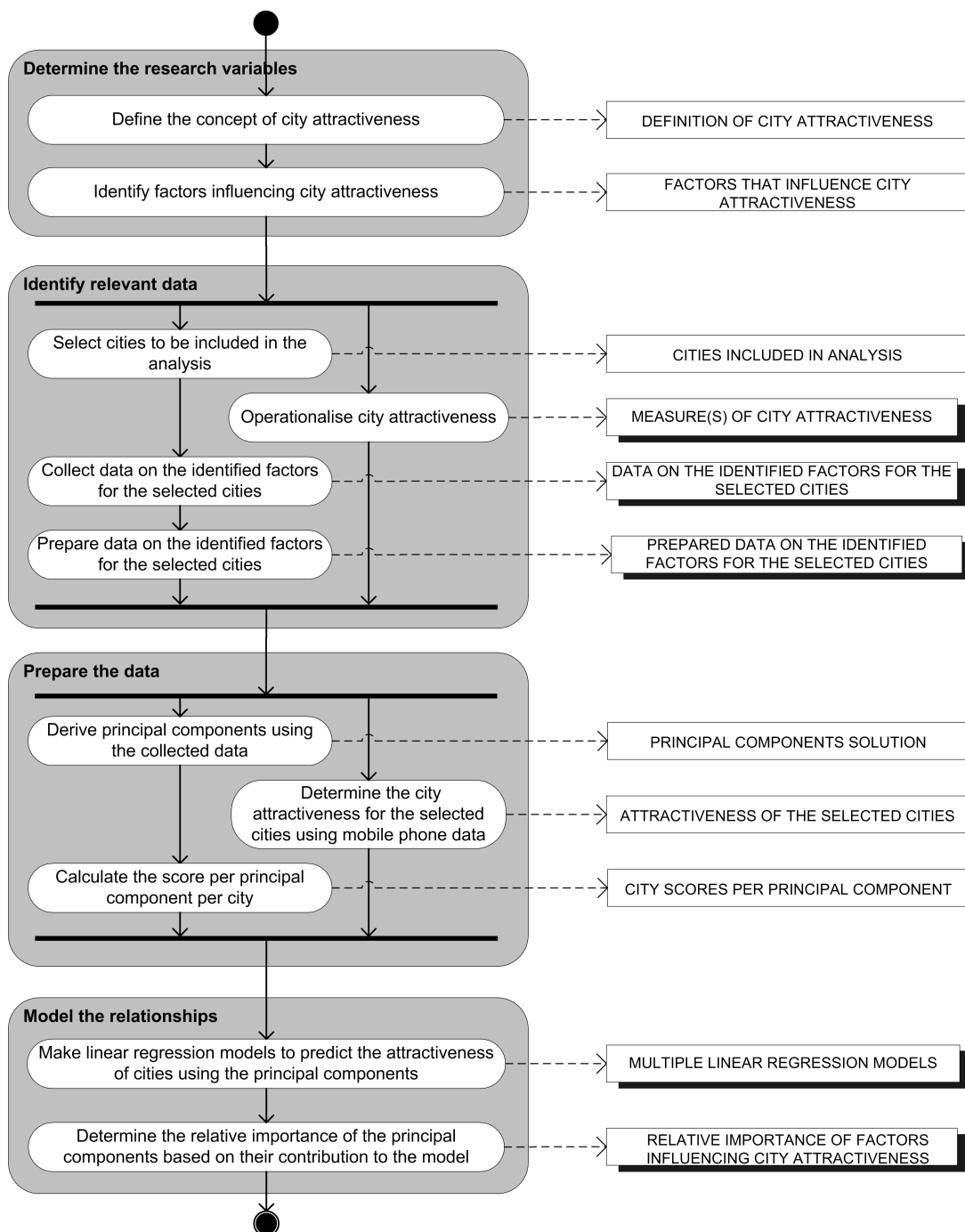


FIGURE 8.6: The proposed method to determine the relative importance of factors that influence city attractiveness.

Chapter 9

Conclusions

9.1 Introduction

This study set out to assess how mobile phone location data can be used to determine the relative importance of factors that influence city attractiveness. Within the context of this research city attractiveness is defined as *the ability of a city to attract visitors and retain residents*, which is based on the definition of city attractiveness by [Servillo et al. \(2011\)](#) who defined city attractiveness as 'the ability to attract or retain specific target groups'.

The benefit of city visitors, i.e. people who live outside the city and come to visit, is that they positively affect income and employments levels in the city ([van der Borg et al., 1996](#)). Therefore, it is important for cities to develop a coherent set of goals and strategies directed towards attracting specific groups of people, rather than to promote at random ([Petrişor-Mateuţ et al., 2013](#)). To attain these goals cities can mobilise assets that can producing changes in the attraction and/or retention of specific segments of the population ([Servillo et al., 2011](#)). However, to support their decisions to mobilise a certain asset, policy makers need information on how certain policy introductions or changes in the built environment, such as the placement of a new mall, will influence the attraction and/or retention of specific segments of the population.

Currently, the options to evaluate these effects are limited to local ad-hoc solutions like visitor counts using wireless signals (e.g., Wi-Fi and Bluetooth) or using cameras (see the interview in [Appendix A](#)). Examples of these techniques can be found in [Li et al. \(2008\)](#); [Weppner and Lukowicz \(2013\)](#); [Xi et al. \(2014\)](#). However, the biggest downside to cameras, Wi-Fi and in particular Bluetooth is the limited range, which is of the order of 10 to 20 meters ([Sadhukhan et](#)

al., 2010). Therefore, the main goal of this study was to assess how mobile phone location data, which does not suffer from these limitations, can be used to determine the relative importance of factors that influence city attractiveness.

To answer this question several sub research questions were formulated, namely: (1) "What is known in literature about the factors that influence city attractiveness?", (2) "What is the quality of the mobile phone location data?", (3) "How can motives for visiting a city be identified from mobile phone location data?".

9.2 Findings

This study has shown a promising and novel approach to use mobile phone location data to empirically determine the relative importance of factors that influence city attractiveness. The method can be used for a variety of research fields including city centre retail attractiveness, urban planning and tourism destination attractiveness to determine the relative importance of factors that influence city attractiveness. The results obtained by applying our method seem logical but cannot be validated using existing literature, because our application was too broad to be able to compare it to earlier studies.

Furthermore, it is very important to understand that the attractiveness of a city differs for various groups within our society (Sinkienė & Kromalcas, 2010). Therefore, this study has developed and implemented a model that can predict the most likely trip motive of trips in the mobile phone location data based on a set of trip attributes such as the arrival time and the day of week. The results of the implemented model are a good match with the trip motive distributions from OViN, i.e. a yearly survey aimed to gain information about mobility patterns of inhabitants of the Netherlands.

Additionally, this study has proposed an improved method to extrapolate the sample, that is present in the mobile phone location data, to the travelling population. The main advantage of the new method is that it takes into account demographic differences of the population, which results in a more accurate representation of the actual population. The results are evaluated to the road-side measurements data by (Meppelink, 2016) and show a correlation ranging from .92 up to .98, depending on the situation.

9.3 Limitations

There are several limitations to the results of this study. These limitations are discussed according to the structure of the report, i.e. starting with the scaling factor, then the trip motives, and finally the proposed method to determine the relative importance of factors influencing city attractiveness using mobile phone location data.

The first limitation of this study applies to the improved method to calculate the scaling factor, introduced in chapter 5. This method contains a calculation for the market share that only takes into account regional differences in the market share. No correction is applied for differences in the market share between age groups. The reason for not including this in the scaling factor was that there was insufficient data available on the market share per age group. If data on the market share per age group does become available and is included, the accuracy of the resulting scaling factor can be improved.

The second limitation applies to the the trip motive prediction model that was developed to determine the most likely trip motive based on a number of trip attributes and trained using [OVIN](#). However, one of the major limitations of [OVIN](#) is that it does not contain a large sample size. Consequently, due to the small sample size of the train set the trip motive prediction model does not take into account many of the important geographical differences between places. This makes it very hard for the model to predict the motive of trips from and to cities that are a little bit out of the ordinary. For example, Almere produces and generates a lot of home-to-work trips and thereby it deviates a lot from the national average distribution of trip motives. Our attempt to artificially include geographical differences (for example by including land use characteristics of the origins and destinations) did not yield good results. In conclusion, the trip motive model works very well at national scales, but the more an specific an area becomes and the more it deviates from the national distribution of trip motives, the harder it becomes to predict the motives of the in- and outbound trips of that area.

The last limitations apply to the way the proposed method has been applied to determine the relative importance of factors influencing city attractiveness using mobile phone location data.

- Although it is generally understand that distance is an important factor for city attractiveness we have not been successful in taking it into account in our analysis (referring to the gravity model discussed in the literature review section 4.5). The problem is that

travel distance is origin-destination specific and our analysis takes place on the level of a city where all in- and outbound trips of a city are aggregated.

- The datasets that have been used to represent the identified factors that influence city attractiveness are not of the same year. This might cause some noise in the results as the values are not compared for the same time period. However, the characteristics represented by the data are not expected to change a lot over time. Hence, this is considered to be a minor limitation.
- The measurement of a city's characteristics using census data is not complete as the perceived image of a city that visitors have can differ from the real image of a city, as discussed by for example [Kalandides, Kavaratzis, and Zenker \(2011\)](#). By taking into account the perceived image by using for example, social media sentiment or city branding budgets or visitor attitudes an interesting and possibly important concept can be captured.
- The influence of weather is not taken into account as a variable. This has its limitations as weather is expected to influence some type of trip motives more than other. For example, [Parsons \(2001\)](#) suggests that the most tangible weather variables (rainfall and temperature) influence shopping destination choice decisions. On the other hand, work-related traffic is expected to be influenced by weather to a lesser degree. Hence, by including weather as a factor it can be expected that additional differences are observed in the relative importance of factors between the trip motive types.

9.4 Future Research

During the research project various ideas have come to our attention. However, many of these ideas could not be researched to their full extent because they were out of the scope of this research project. Therefore, in this section, an outline is given of these ideas to serve as a source of inspiration for future research.

First, it would be very interesting to expand the segmentation of travel motive, described in chapter 6, to other dimensions. Perhaps the user's mobility pattern can tell a lot more about the person than just the motive for performing this travel and it can be used to determine users' gender or age. Additional characteristics could be very useful in feasibility studies and marketing campaigns.

Secondly, the use of mobile phone location data as a data source on its own is not enough to apply to all target groups, which a city potentially wishes to attract (see (Braun, 2008)). Obviously, the attraction of permanent migration, of for example manufacturers, was outside of the scope of this research. However, cities are not only interested in what influences visitors to come to their city, but also in attracting corporate headquarters, manufacturers and exporters (Kotler, 2002). For example, combining the results of a survey with the mobile phone location data could provide further insight into the perceived image of city that people have. However, the number of respondents of a surveys is almost linear proportional to the cost of the survey, which often results in a limited amount of respondents. Thereby, important geographical differences cannot be represented by survey data.

Thirdly, the proposed method can be applied in a longitudinal study using mobile phone location data spanning one or multiple years. This way trends in changes in the relative importance of the determinants of a city's attractiveness could be identified.

Finally, Teller and Reutterer (2008) studied the relative importance of exogenous factors, such as parking facilities and retail-tenant mix, that influence the perceived city centre retail attractiveness. Their research lacks external validity as they only studied people in two different shopping areas. Consequently, in their recommendations for future research they mention that future research can focus more on benchmarking and inter-city comparison. One of the major advantages of using mobile phone location data for this type of research is the ability for inter-city comparisons. Therefore, it would be very interesting to see if the proposed method to determine the relative importance of factors that influence city attractiveness can successfully be applied to analyse city centres.

References

- Abdi, H. (2003). Factor rotations in factor analyses. *Encyclopedia for Research Methods for the Social Sciences*. Sage: Thousand Oaks, CA, 792–795.
- Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An estonian case study. *Tourism Management*, 29(3), 469–486.
- A.T. Kearney's Global Cities Index. (2015). <https://www.atkearney.com/research-studies/global-cities-index/2015>. (Retrieved: 2015-07-31)
- Baker, R. G. (2006). *Dynamic trip modelling*. Springer.
- Becker, R. A., Cáceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A., & Volinsky, C. (2011). A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4), 18–26.
- Berli, A., & Martin, J. D. (2004). Factors influencing destination image. *Annals of tourism research*, 31(3), 657–681.
- Begg, I. (1999). Cities and competitiveness. *Urban studies*, 36(5-6), 795–809.
- Borgegård, L.-E., & Murdie, R. (1993). Socio-demographic impacts of economic restructuring on stockholm's inner city. *Tijdschrift voor economische en sociale geografie*, 84(4), 269–280.
- Braun, E. (2008). *City marketing: Towards an integrated approach*. Erasmus Research Institute of Management (ERIM).
- Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5), 1287–1294.
- Brounen, D., Kok, N., & Quigley, J. M. (2012). Residential energy use and conservation: Economics and demographics. *European Economic Review*, 56(5), 931–945.
- Brouwer, F. (2015). *Applying Semantic Integration to improve Data Quality* (Unpublished master's thesis). Utrecht University, Utrecht, The Netherlands.

- Bruneckiene, J., Guzavicius, A., & Cincikaite, R. (2010). Measurement of urban competitiveness in lithuania. *Engineering Economics*, 21(5), 493–508.
- Bruton, M. J. (1985). Introduction to transportation planning.
- Burns, D. J., & Warren, H. B. (1995). Need for uniqueness: shopping mall preference and choice activity. *International Journal of Retail & Distribution Management*, 23(12), 4–12.
- Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning* (pp. 161–168).
- CBS. (2014). *Onderzoek Verplaatsingen in Nederland 2014 - OVIN 2014*. DANS. <http://dx.doi.org/10.17026/dans-x95-5p7y>. (Retrieved: 2015-10-02)
- CBS. (2015). *Onderzoek Verplaatsingen in Nederland 2014 Plausibiliteitsanalyse*. <http://www.cbs.nl/NR/rdonlyres/ECD2998A-D5A2-4599-AF46-BA77495ED071/0/2015OVINplausibiliteitsanalysepub.pdf>. (Retrieved: 2016-01-02)
- Chambers, J. M. (1991). Linear models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical models in s*. Boca Raton, FL, USA: CRC Press, Inc.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.
- Clark, T. N., Lloyd, R., Wong, K. K., & Jain, P. (2002). Amenities drive urban growth. *Journal of urban affairs*, 24(5), 493–515.
- Cohen, B. (2015). Urbanization, city growth, and the new united nations development agenda. *Cornerstone, The Official Journal of the World Coal Industry*, 3(2), 4–7.
- Cussens, J. (1993). Bayes and pseudo-bayes estimates of conditional probabilities and their reliability. In *Machine learning: Ecml-93* (pp. 136–152).
- Daas, P., Roos, M., de Blois, C., Hoekstra, R., ten Bosch, O., & Ma, Y. (2011). *New data sources for statistics: experiences at statistics netherlands*. Heerlen/Den Haag, The Netherlands: Statistics Netherlands.
- Deas, I., & Giordano, B. (2001). Conceptualising and measuring urban competitiveness in major english cities: an exploratory approach. *Environment and Planning A*, 33(8), 1411–1430.
- Delbari, S. A., Ng, S. I., Aziz, Y. A., & Ho, J. A. (2015). Measuring the influence and impact of competitiveness research: a web of science approach. *Scientometrics*, 1–16.
- de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3.
- Dennis, C. (2005). *Objects of desire: Consumer behaviour in shopping centre choices*. Palgrave Macmillan.

- Dietterich, T. G. (1996). Statistical tests for comparing supervised classification learning algorithms. In (Vol. 10, pp. 1895–1923).
- Eagle, N., Pentland, A. S., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, *106*(36), 15274–15278.
- Enright, M. J., & Newton, J. (2004). Tourism destination competitiveness: a quantitative approach. *Tourism management*, *25*(6), 777–788.
- European Commission. (2000). *European competitiveness report 2000: Working document of the services of the european commission*. Office for official publications of the European Communities.
- Farag, S., Schwanen, T., Dijst, M., & Faber, J. (2007). Shopping online and/or in-store? a structural equation model of the relationships between e-shopping and in-store shopping. *Transportation Research Part A: Policy and Practice*, *41*(2), 125–141.
- Ferri, C., Flach, P. A., & Hernández-Orallo, J. (2003). Improving the auc of probabilistic estimation trees. In *Machine learning: Ecml 2003* (pp. 121–132). Springer.
- Field, A. (2009). *Discovering statistics using spss*. Sage publications.
- Formica, S., & Uysal, M. (2006). Destination attractiveness based on supply and demand evaluations: An analytical framework. *Journal of Travel Research*, *44*(4), 418–430.
- Geerts, J. (2014). *Ophogingsmethodiek van devices naar personen*. (Unpublished internal document)
- Glaeser, E. L. (1998). Are cities dying? *Journal of Economic Perspectives*, *12*(2), 139–160.
- Global Power City Index. (2014). http://www.mori-m-foundation.or.jp/gpci/index_e.html. (Retrieved: 2015-07-31)
- Grömping, U., et al. (2006). Relative importance for linear regression in r: the package relaimpo. *Journal of statistical software*, *17*(1), 1–27.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L., et al. (2006). *Multivariate data analysis*. Pearson Prentice Hall Upper Saddle River, NJ.
- Healey, M. J., & Dunham, P. J. (1994). Changing competitive advantage in a local economy: the case of coventry, 1971-90. *Urban Studies*, *31*(8), 1279–1301.
- Hijmans, R. J. (2015). geosphere: Spherical trigonometry [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=geosphere> (R package version 1.5-1)
- Hofmann, R. J. (1978). Complexity and simplicity as objective indices descriptive of factor solutions. *Multivariate Behavioral Research*, *13*(2), 247–250.

- Jansen-Verbeke, M. (1988). Leisure, recreation and tourism in inner cities. explorative case-studies. *Nederlandse Geografische Studies*(58).
- Kalandides, A., Kavaratzis, M., & Zenker, S. (2011). How to catch a city? the concept and measurement of place brands. *Journal of Place Management and Development*, 4(1), 40–52.
- Keij, J. (2014). *Smart phone counting: Location-Based Applications using mobile phone location data* (Unpublished master's thesis). Delft University of Technology, Delft, The Netherlands.
- Khei Mie Wong, G., Lu, Y., & Lan Yuan, L. (2001). Scattr: an instrument for measuring shopping centre attractiveness. *International Journal of Retail & Distribution Management*, 29(2), 76–86.
- Kitson, M., Martin, R., & Tyler, P. (2004). Regional competitiveness: an elusive yet key concept? *Regional studies*, 38(9), 991–999.
- Kitson, M., Martin, R., & Tyler, P. (2005). The regional competitiveness debate.
- Kotler, P. (2002). *Marketing places*. Simon and Schuster.
- Krešić, D., & Prebežac, D. (2011). Index of destination attractiveness as a tool for destination attractiveness assessment. *Turizam: znanstveno-stručni časopis*, 59(4), 497–517.
- Kresl, P. K. (1995). The determinants of urban competitiveness. In P. K. Kresl & G. Gappert (Eds.), *North american cities and the global economy* (pp. 45–68). Sage, London.
- Lamport, L. (1979). *Constructing digital signatures from a one-way function* (Tech. Rep.). Technical Report CSL-98, SRI International Palo Alto.
- Levinson, D., & Kumar, A. (1995). *A multi-modal trip distribution model* (Working Papers No. 199503). University of Minnesota: Nexus Research Group.
- Li, M., Zhang, Z., Huang, K., & Tan, T. (2008). Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *Pattern recognition, 2008. icpr 2008. 19th international conference on* (pp. 1–4).
- Lidwell, W., Holden, K., & Butler, J. (2010). *Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design*. Rockport Pub.
- Malecki, E. J. (2000). Knowledge and regional competitiveness (wissen und regionale wettbewerbsfähigkeit). *Erdkunde*, 334–351.
- Malecki, E. J. (2002). Hard and soft networks for urban competitiveness. *Urban Studies*, 39(5-6), 929–945.

- Marbán, Ó., Mariscal, G., & Segovia, J. (2009). A data mining & knowledge discovery process model. *Data Mining and Knowledge Discovery in Real Life Applications. IN-TECH, 2009*, 8.
- Margineantu, D. D., & Dietterich, T. G. (2003). Improved class probability estimates from decision tree models. In *Nonlinear estimation and classification* (pp. 173–188). Springer.
- Martin, W. A., McGuckin, N. A., McGuckin, N. A., & McGuckin, N. A. (1998). *Travel estimation techniques for urban planning* (Vol. 365). National Academy Press Washington, DC.
- Meppelink, J. (2016). *Evaluating and Predicting the Impact of Roadworks Using Mobile Phone Movement Data* (Unpublished master's thesis). Utrecht University, Utrecht, The Netherlands.
- Murillo, J., Vayá, E., Romani, J., & Suriñach, J. (2013). How important to a city are tourists and day-trippers? the economic impact of tourism on the city of barcelona. *Tourism Economics, 19*(4), 897–917.
- Ni, P. (2012). *The global urban competitiveness report - 2011*. Edward Elgar Publishing.
- Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on machine learning* (pp. 625–632).
- Ofcom. (2014, October). *Children and parents: Media use and attitudes report 2014* (Tech. Rep.). Southwark, London.
- Offermans, M., Priem, A., & Tennekes, M. (2013). Rapportage project impact ict; mobiele telefonie. *Programma Impact ICT, 1*(9).
- Öner, Ö. (2015). *Retail city: The relationship between place attractiveness and accessibility to shops* (Working Paper Series No. 1055). Research Institute of Industrial Economics.
- Palchykov, V., Kaski, K., Kertész, J., Barabási, A.-L., & Dunbar, R. I. (2012). Sex differences in intimate relationships. *Scientific reports, 2*.
- Pampel, H., Pfeifferberger, H., Schäfer, A., Smit, E., Pröll, S., & Bruch, C. (2012). Report on peer review of research data in scholarly communication. *APARSEN*.
- Parsons, A. G. (2001). The association between daily weather and daily shopping patterns. *Australasian Marketing Journal (AMJ), 9*(2), 78–84.
- Petersen, D. C. (2005). The city as a destination. *Journal of Convention & Event Tourism, 6*(1-2), 145-157.
- Petrișor-Mateuț, O., Orboi, D., & Popa, M. (2013). Management strategies on the promotion of world cities from theory to practice. *Lucrări Științifice: Management Agricol, 15*(2),

- 201–204.
- Pike, S. (2002). Destination image analysis—a review of 142 papers from 1973 to 2000. *Tourism management*, *23*(5), 541–549.
- Porter, M. E. (1990). The competitive advantage of nations. *Harvard business review*, *68*(2), 73–93.
- Postcode Data. (2014). *Postcode informatie van Nederland*. <http://www.postcodedata.nl/download/>. (Retrieved: 2015-12-08)
- Prettig Parkeren. (2015). *Parkeertarieven*. <https://www.prettigparkeren.nl/>. (Retrieved: 2015-12-20)
- Provost, F., & Domingos, P. (2000). Well-trained pets: Improving probability estimation trees.
- Provost, F., & Domingos, P. (2003). Tree induction for probability-based ranking. *Machine Learning*, *52*(3), 199–215.
- R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <http://www.R-project.org/>
- Richards, J. W., Starr, D. L., Miller, A. A., Bloom, J. S., Butler, N. R., Brink, H., & Crellin-Quick, A. (2012). Construction of a calibrated probabilistic classification catalog: Application to 50k variable sources in the all-sky automated survey. *The Astrophysical Journal Supplement Series*, *203*(2), 32.
- Rüping, S. (2006). Robust probabilistic calibration. In *Machine learning: Ecml 2006* (pp. 743–750). Springer.
- Sadhukhan, P., Chatterjee, N., Das, A., & Das, P. K. (2010). A scalable location-based services infrastructure combining gps and bluetooth based positioning for providing services in ubiquitous environment. In *Internet multimedia services architecture and application (imsaa), 2010 ieee 4th international conference on* (pp. 1–6).
- Servillo, L., Atkinson, R., & Russo, A. P. (2011). Territorial attractiveness in eu urban and spatial policy: A critical review and future research agenda. *European Urban and Regional Studies*, *19*(4), 349–365.
- Singhal, S., McGreal, S., & Berry, J. (2013). Application of a hierarchical model for city competitiveness in cities of india. *Cities*, *31*, 114–122.
- Sinkienė, J. (2008). Factors of urban competitiveness. *Public Policy and Administration*, *1*(25).
- Sinkienė, J. (2009). Competitiveness factors of cities in lithuania. *Public Policy and Administration*, *1*(29).
- Sinkienė, J., & Kromalcas, S. (2010). Concept, directions and practice of city attractiveness improvement. *Public Policy and Administration*, *1*(31), 147–154.

- Snijkers, G. (2009). Getting data for (business) statistics: What's new? what's next. In *European conference for new techniques and technologies for statistics (ntts)* (pp. 18–20).
- Swarbrooke, J., & Page, S. J. (2012). *Development and management of visitor attractions*. Routledge.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Allyn and Bacon Boston.
- Telecompaper. (2015, Q3). *Smartphone penetration netherlands 2015 q3* (Tech. Rep.). Southwark, London.
- Teller, C., & Reutterer, T. (2008). The evolving concept of retail attractiveness: what makes retail agglomerations attractive when customers shop at them? *Journal of Retailing and Consumer Services*, 15(3), 127–143.
- Therneau, T., Atkinson, B., & Ripley, B. (2014). rpart: Recursive partitioning and regression trees [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=rpart> (R package version 4.1-8)
- Thurstone, L. L. (1935). The vectors of mind: Multiple-factor analysis for the isolation of primary traits.
- Ullman, E. L. (1954). Amenities as a factor in regional growth. *Geographical Review*, 44(1), 119–132.
- United Nations. (2014). World urbanization prospects: The 2014 revision. *United Nations, Department of Economic and Social Affairs (DESA), Population Division, Population Estimates and Projections Section, New York*.
- van Bergeijk, P. A., & Brakman, S. (2010). *The gravity model in international trade: Advances and applications*. Cambridge University Press.
- van der Borg, J., Costa, P., & Gotti, G. (1996). Tourism in european heritage cities. *Annals of Tourism Research*, 23(2), 306–321.
- van de Weerd, I., & Brinkkemper, S. (2009). Meta-modeling for situational analysis and design methods. In M. R. Syed & N. Syed (Eds.), *Handbook of research on modern systems analysis and design technologies and applications* (pp. 35–54). IGI Global.
- Van Kats, J. (2014). *Supporting Incident Management with Population Information Derived from Telecom Data* (Unpublished master's thesis). Utrecht University, Utrecht, The Netherlands.
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3), 321–327.

- Weltevreden, J. W. J. (2006). City centres in the internet age: Exploring the implications of b2c e-commerce for retailing at city centres in the netherlands.
- Weltevreden, J. W. J., & van Rietbergen, T. (2007). E-shopping versus city centre shopping: The role of perceived city centre attractiveness. *Tijdschrift voor economische en sociale geografie*, 98(1), 68–85.
- Weppner, J., & Lukowicz, P. (2013). Bluetooth based collaborative crowd density estimation with mobile phones. In *Pervasive computing and communications (percom), 2013 ieee international conference on* (pp. 193–200).
- Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(1), 1–23.
- Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29–39).
- Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering* (p. 38).
- Wrigley, N., & Lowe, M. (2002). Reading retail: A geographical perspective on retailing and consumption spaces.
- Xi, W., Zhao, J., Li, X.-Y., Zhao, K., Tang, S., Liu, X., & Jiang, Z. (2014). Electronic frog eye: Counting crowd using wifi. In *Infocom, 2014 proceedings ieee* (pp. 361–369).
- Zadrozny, B., & Elkan, C. (2001). Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the seventh acm sigkdd international conference on knowledge discovery and data mining* (pp. 204–213).
- Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining* (pp. 694–699).
- Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, 99(3), 432.

Appendix A

Interview on City Attractiveness

Please note that this interview is written, conducted and reported in Dutch as all participants are expected to prefer to speak Dutch. Moreover, by conversing in their native tongue they will not be limited in expressing their thoughts and answering the questions.

Introductie

Jens: In mijn onderzoek richt ik mij op de vraag of de mobiele telefonie data van Mezero gebruikt kan worden voor het bepalen van de aantrekkingskracht van steden voor bezoekers.

Hoe heet u? Hans Brouwers

Waar werkt u? Centrummanagement Sittard

Wat is uw functie? Centrummanager voor Sittard

Wat zijn uw voornaamste werkzaamheden?

Hans: Op dit moment reorganiseren, maar daarvoor bestonden de werkzaamheden uit promoten, organiseren van evenementen, contacten onderhouden met ondernemers en overheid, beleid uitstippelen, en innoveren.

Een Definitie

Hoe zou u de aantrekkingskracht van een stad definiëren? Mijn definitie is als volgt: De aantrekkingskracht van een stad is de mate waarin een stad mobiliteitsstromen richting de stad weet te realiseren.

Hans: Het is belangrijk dat je de bezoekers een compleet pakket aan biedt. Zakelijk moet je goederen en diensten aanbieden, als het om winkelen gaat moet je een divers aanbod hebben, goede horeca en het in het algemeen er netjes uit ziet. Daarnaast moeten de basiszaken op orde zijn, zoals de infrastructuur, parkeren, bewegwijzering. In Roermond hebben ze naar het voorbeeld van interconnect birmingham de bewegwijzering veranderd. Als een bezoeker een keuze moet maken om een stad te bezoeken die zich binnen een half uur reizen bevindt dan moet het complete plaatje van de stad kloppen om de bezoeker te kunnen aantrekken. Mensen die Roermond bezoeken geven aan dat ze de stad een verrassend mooie stad vinden met veel oude gebouwen. Als het plaatje klopt dan maakt 75% de keuze om die stad te bezoeken.

Algemeen

Welke informatiebronnen raadpleegt u tijdens het uitvoeren van uw werkzaamheden? Bijvoorbeeld externe bronnen zoals het CBS of Google, of interne bronnen zoals Google Drive, Sharepoint of collega's.

Hans: Wifi punten tellen hoeveel mensen er langs komen in Roermond. Daarmee kan je zien hoe bezoekers bewegen in de stad. waar komen ze vandaan, wat ze besteden, of ze terug komen, en hoe lang ze blijven. Winkels hebben scanners staan aan de ingang van de zaak die het aantal bezoekers tellen. Outlets weten waar mensen vandaan komen op basis van de transacties die mensen verrichten in de winkels. Winkels geven deze informatie door aan de de beheerder van het outlet center. Daarmee kan marketing gericht worden uitgevoerd.

Bij welke werkzaamheden gebruikt u de aantrekkingskracht van een stad? Hans: De ontwikkelingen in de outlet center worden gedeeld met de city marketing Sittard. De Rabobank heeft ook baat bij deze informatie omdat deze de winkels financiert. Ze zijn gebaat bij een goede economische regio. Het belang van de bank is ontzettend groot.

Jens: Welke groep mensen is belangrijk voor het in stand houden van de stad?

Hans: De consument die binnen een straal van 20km komt is verantwoordelijk voor 75

Jens: Is het slim om je te richten op het aantrekken van forenzen?

Hans: Nee, die hebben een andere beeld van de stad waardoor je daar echt op zou moeten richten. Je moet je richten op recreanten. Alles binnen een drie kwartier tot een uur rijden van de stad is het gros van de mensen die naar de stad toe komen (20km uit onderzoek van de Rabobank en Mezero).

Kunt u een voorbeeld geven van de informatie die u op zoekt?

Hans: bestedingsinformatie en het samenvoegen van informatie.

De aantrekkingskracht van een stad kan in verschillende doelgroepen worden opgesplitst. Denk aan forenzen, winkelend publiek, dagjesmensen, toeristen. Is er een mobiliteitsstroom die het belangrijkste is? En waarom denk u dat?

Hans: De aanpassing aan de bewegwijzering zorgt waarschijnlijk voor een betere penetratiegraad van bezoekers die naar het centrum gaan.

Stel de informatievoorziening zou perfect zijn (u zou alle informatie kunnen krijgen die u maar wil), welke informatie zou u dan willen hebben om te bepalen wat de aantrekkelijkheid van uw stad is?

Hans: Steden die hun winkelaanbod hebben aangesloten op de wensen van de bezoeker hebben een voorsprong op steden die dat niet hebben gedaan. Mensen kiezen voor een stad die hun bekend is en de zaken op orde heeft. Dus verkeersproblematiek kan er zeker voor zorgen dat er minder bezoekers naar een stad toe komen.

Jens: Hoe komt u achter gegevens of u bezoekers van andere steden terug pakt en dus eigenlijk concurrerend bezig bent?

Hans: Dat is heel moeilijk. Het kan hoogstens gerelateerd worden aan dingen als leegstand. Met de gegevens van Mezero is dit al beter mogelijk. Maar het kunnen sturen op concrete gegevens is iets wat steden missen.

Jens: Wie zijn nou de bezoekers met de meeste koopkracht? Hans: De Rabobank publiceert iedere 4 jaar een monitor waarin gegevens staan over bezoekers. De Rabobank financiert ook de SWOT analyses en onderzoek voor beginnende bedrijven.

Uw stad

Vergelijkt u “uw stad” met andere steden in de zin van aantrekkelijkheid?

Hans: Nee, dat kan niet. Maar op sommige punten kan dat wel bijvoorbeeld Roermond heeft 2009-2011 de beste binnenstad. Als je het goed zou willen doen zou je een enquête moeten houden onder de inwoners van de gemeente Limburg.

Zoekt u ook wel eens informatie over bewegingsstromen van mensen die naar uw stad komen?

Hans: Nee, dat is heel moeilijk. Elsevier doet dat wel, die heeft een bepaalde attractiviteitsmeter. Dat is een monitor die op heel veel verschillende aspecten meet.

Heeft u nog contacten die u zou willen aanraden?

Hans: binnenstadmanagement, boxtel bureau trommelen, adviesbureau nbs, nieuwe binnenstadmanagement shopping2020. bro binnenstadmanagement

Appendix B

Proposed Method

The method depicted in figure [B.1](#) is the final deliverable of this study. The activity table can be found in table [B.1](#). The concept table is presented in table [B.2](#).

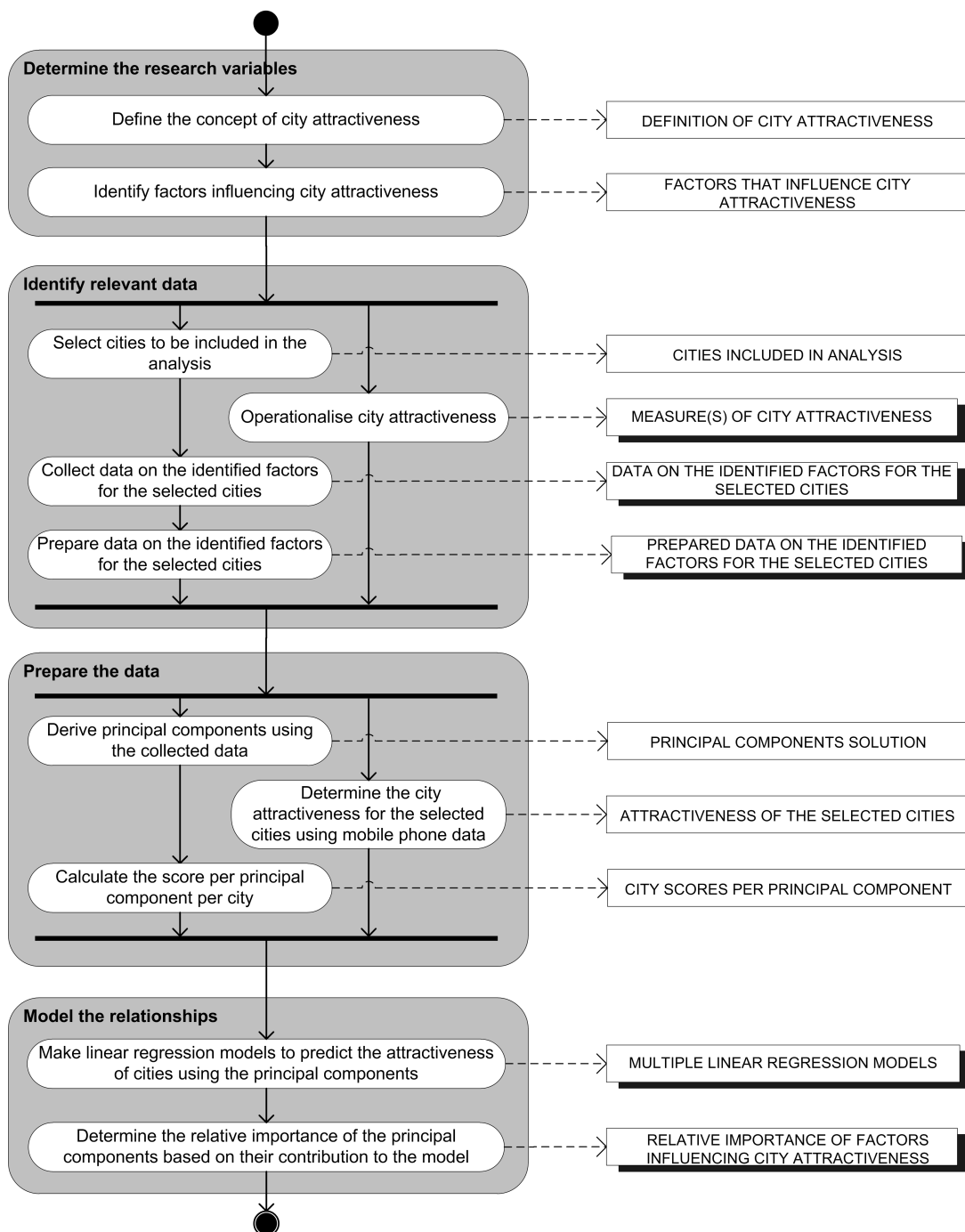


FIGURE B.1: The proposed method to determine the relative importance of factors that influence city attractiveness.

TABLE B.1: Descriptions of the activities and sub-activities that are used in the method depicted in figure B.1

Activity	Sub-activity	Description
Determine the re-search variables	Define the concept of city attractiveness	Determine a definition for the concept of city attractiveness useful for the research.
	Identify factors influencing city attractiveness	Use existing literature to determine which factors influence city attractiveness according to your DEFINITION OF CITY ATTRACTIVENESS.
Identify relevant data	Select cities to be included in the analysis	Determine which cities are of interest to the research.
	Collect data on the identified factors for the selected cities	Collect data on the FACTORS THAT INFLUENCE CITY ATTRACTIVENESS for each of the cities in the CITIES INCLUDED IN ANALYSIS.
	Prepare data on the identified factors for the selected cities	Organise the data in a table with the cities as rows and factors as columns according to Wickham's (2014) tidy data principles
	Operationalise city attractiveness	Determine how city attractiveness is calculated using mobile phone location data, while satisfying the DEFINITION OF CITY ATTRACTIVENESS.
Prepare the data	Derive principal components using the collected data	Use principal components analysis to derive principal components (i.e., composite variables) from the DATA ON THE IDENTIFIED FACTORS FOR THE SELECTED CITIES. Make sure to test the assumptions of a principal component analysis.
	Calculate the score per principal component per city	Use multiple regression to determine the scores per city per principal component. The variable loadings from the PRINCIPAL COMPONENTS SOLUTION are the independent variables and the DATA ON THE IDENTIFIED FACTORS FOR THE SELECTED CITIES are the coefficients (see Thurstone, 1935).
	Determine the city attractiveness for the selected cities using mobile phone data	Using the mobile phone location data determine for each city the MEASURE(S) OF CITY ATTRACTIVENESS.
Model the relationship	Make linear regression models to predict the attractiveness of the selected cities using the principal components	Use multiple regression analysis to build regression models for each MEASURE(S) OF CITY ATTRACTIVENESS. The MEASURE(S) OF CITY ATTRACTIVENESS serves as dependent variable and the principal components from the PRINCIPAL COMPONENTS SOLUTION serve as independent variables.

Continued on the next page

TABLE B.2: Descriptions of the concepts that are used in the method depicted in figure B.1

Concept	Description
DEFINITION OF CITY ATTRACTIVENESS	A definition of city attractiveness.
FACTORS THAT INFLUENCE CITY ATTRACTIVENESS	A set of factors identified from literature that are known to influence city attractiveness as defined in the DEFINITION OF CITY ATTRACTIVENESS.
CITIES INCLUDED IN ANALYSIS	A list of cities that are selected to be included within this research.
MEASURE(S) OF CITY ATTRACTIVENESS	One or more formulas that specify how city attractiveness is calculated, including a description and preferably examples.
DATA ON THE IDENTIFIED FACTORS FOR THE SELECTED CITIES	A dataset that contains information on each of the identified FACTORS THAT INFLUENCE CITY ATTRACTIVENESS for each of the cities in the CITIES INCLUDED IN ANALYSIS.
PREPARED DATA ON THE IDENTIFIED FACTORS FOR THE SELECTED CITIES	This is the DATA ON THE IDENTIFIED FACTORS FOR THE SELECTED CITIES structured and ready for statistical analyses.
ATTRACTIVENESS OF THE SELECTED CITIES	The data from the mobile phone location data obtained according to the MEASURE(S) OF CITY ATTRACTIVENESS for each of the CITIES INCLUDED IN ANALYSIS
PRINCIPAL COMPONENTS SOLUTION	The result of the principal component analysis, which are the loadings of each variable on each of the principal components.
CITY SCORES PER PRINCIPAL COMPONENT	A table of values the size of all cities time the number of factors. Each field contains the score of the respective city on that principal component
MULTIPLE LINEAR REGRESSION MODELS	A set of linear regression models. One for each MEASURE(S) OF CITY ATTRACTIVENESS in which the principal components are the independent variables
RELATIVE IMPORTANCE OF FACTORS INFLUENCING CITY ATTRACTIVENESS	The relative distribution of each of the dependent variables from the MULTIPLE LINEAR REGRESSION MODEL in terms of its contribution to the overall R^2 . (see Grömping, 2006).

TABLE B.1: (continued)

Activity	Sub-activity	Description
	Determine the relative importance of the principal components based on their contribution to the model	Determine the relative importance of the principal components based on their contribution to the overall R^2 of the model (see Grömping, 2006).

Appendix C

Motive Prediction in OViN using R

Every year [CBS](#) performs a large survey called [OViN](#) in which they ask members of the Dutch population about their travel patterns. On average, 35,000 people are asked to fill in a travel diary for a single day in their life. In this research we use the results from this survey to build a model that can be used to predict the trip motive on some known attributes of a trip. Some of the attributes in the mobile phone location data are not available in [OViN](#). How these attributes are computed is shown in the code below, which is written in the R language. This serves as a reference to determine how the attributes are calculated.

Add Characteristics to OViN

```
1
2 #-----
3 # Description: Deduce trip and person characteristics that correspond to the
4               data available in the mobile phone location data to build the model
5 # Input:      modified ovin 2013 dataset
6 # Output:     ovin 2013 dataset greater than 10 km
7 #-----
8 # Select the trips from OViN greater than 10 km
9 ovin_over_ten_km <- ovin[as.integer(ovिन$KAFSTV) > 8,]
10 # 8 = "7,5 tot 10 km"
11
12 # Select the trips from OViN with a duration of stay at the destination greater
    than 30 minutes
```

```

13 ovin_over_ten_km <- ovin_over_ten_km[ ovin_over_ten_km$ACTDUUR > 30 ,]
14 # Exclude empty rows
15 ovin_over_ten_km <- ovin_over_ten_km[!is.na( ovin_over_ten_km$KAFSTV) ,]
16
17 # Convert hours to minutes
18 ovin_over_ten_km$VERTREKTIJD <- ovin_over_ten_km$VERTUUR * 60 + ovin_over_ten_km$
  VERTMIN
19 ovin_over_ten_km$AANKOMSTTIJD <- ovin_over_ten_km$AANKUUR * 60 + ovin_over_ten_km
  $AANKMIN
20
21 # Set trip distance data type as ordinal
22 ovin_over_ten_km$KAFSTV <- ordered( ovin_over_ten_km$KAFSTV)
23
24 # Add a field to each trip , indicating whether it is towards home
25 ovin_over_ten_km$DESTHOME[ ovin_over_ten_km$DOEL == "Naar huis" ] <- 1
26 ovin_over_ten_km$DESTHOME[ ovin_over_ten_km$DOEL != "Naar huis" ] <- 0
27
28 # Make the transportation modes correspond to the ones available in the mobile
  phone location data
29 ovin_over_ten_km$MODALITY <- "Other"
30 ovin_over_ten_km$MODALITY[ ovin_over_ten_km$KHVM == "Auto als bestuurder" ] <- "
  Vehicle"
31 ovin_over_ten_km$MODALITY[ ovin_over_ten_km$KHVM == "Auto als passagier" ] <- "
  Vehicle"
32 ovin_over_ten_km$MODALITY[ ovin_over_ten_km$KHVM == "Trein" ] <- "Train"
33 ovin_over_ten_km$TRAIN[ ovin_over_ten_km$MODALITY == "Train" ] <- 1
34 ovin_over_ten_km$TRAIN[ ovin_over_ten_km$MODALITY != "Train" ] <- 0
35 ovin_over_ten_km$VEHICLE[ ovin_over_ten_km$MODALITY == "Vehicle" ] <- 1
36 ovin_over_ten_km$VEHICLE[ ovin_over_ten_km$MODALITY != "Vehicle" ] <- 0
37
38 # Calculate the number of trips per respondent per day
39 total.trips <- aggregate( ovin_over_ten_km$OPID, by = list( ovin_over_ten_km$OPID),
  length)
40 ovin_over_ten_km$TRIP_FREQUENCY <- merge( ovin_over_ten_km, total.trips , by.x = "
  OPID" , by.y = "Group.1" , all = T)$x
41
42 # Calculate the number of trips per respondent per day during daytime (6h – 19h)
43 temp <- ovin_over_ten_km[ ovin_over_ten_km$VERTUUR >= 6 ,]

```

```

44 temp <- temp[temp$VERTUUR <= 19,]
45 total.trips.day <- aggregate(temp$OPID, by = list(temp$OPID), length)
46 rm(temp)
47 ovin_over_ten_km$TRIP_FREQUENCY_DAY <- merge(ovin_over_ten_km, total.trips.day,
      by.x = "OPID", by.y = "Group.1", all= T)$x
48
49 # Number of trips per respondent per day
50 total.trips.day.base <- as.data.frame(total.trips$Group.1)
51 total.trips.day <- merge(total.trips.day, total.trips.day.base, by.x = "Group.1",
      by.y = "total.trips$Group.1", all = T)
52 rm(total.trips.day.base)
53 total.trips.day[is.na(total.trips.day$x),]$x <- 0
54
55 # Departure time of the first trip
56 first.trip.start <- aggregate(ovin_over_ten_km$VERTREKTIJD, by = list(ovin_over_
      ten_km$OPID), min)
57 ovin_over_ten_km$FIRST_TRIP_START <- merge(ovin_over_ten_km, first.trip.start,
      by.x = "OPID", by.y = "Group.1", all = T)$x
58 # Arrival time of the first trip
59 first.trip.end <- aggregate(ovin_over_ten_km$AANKOMSTTIJD, by = list(ovin_over_
      ten_km$OPID), min)
60 ovin_over_ten_km$FIRST_TRIP_END <- merge(ovin_over_ten_km, first.trip.end, by.x =
      "OPID", by.y = "Group.1", all = T)$x
61 # Departure time of the last trip
62 last.trip.start <- aggregate(ovin_over_ten_km$VERTREKTIJD, by = list(ovin_over_
      ten_km$OPID), max)
63 ovin_over_ten_km$LAST_TRIP_START <- merge(ovin_over_ten_km, last.trip.start, by.x
      = "OPID", by.y = "Group.1", all = T)$x
64 # Arrival time of the last trip
65 last.trip.end <- aggregate(ovin_over_ten_km$AANKOMSTTIJD, by = list(ovin_over_ten
      km$OPID), max)
66 ovin_over_ten_km$LAST_TRIP_END <- merge(ovin_over_ten_km, last.trip.end, by.x = "
      OPID", by.y = "Group.1", all = T)$x
67
68 # Trip taken during weekend
69 ovin_over_ten_km$WEEKEND[ovin_over_ten_km$WEEKDAG == "Zaterdag"] <- 1
70 ovin_over_ten_km$WEEKEND[ovin_over_ten_km$WEEKDAG == "Zondag"] <- 1
71 ovin_over_ten_km$WEEKEND[is.na(ovin_over_ten_km$WEEKEND)] <- 0

```

```

72
73 # The following attributes are at the level of the respondent instead of the
      level of a trip
74 # Respondent attribute: Traveled during the weekend
75 weekend <- aggregate(ovin_over_ten_km$WEEKEND, by = list(ovin_over_ten_km$OPID),
      max)
76 # Respondent attribute: Total travel time
77 total.travel.time <- aggregate(ovin_over_ten_km$AANKOMSTTIJD - ovin_over_ten_km$
      VERTREKTIJD, by = list(ovin_over_ten_km$OPID), sum)
78 ovin_over_ten_km$TIME_TRIPPING <- merge(ovin_over_ten_km, total.travel.time, by.x
      = "OPID", by.y = "Group.1", all = T)$x
79
80 # Respondent attribute: Trips not to or from home
81 trips.not.home <- aggregate(ovin_over_ten_km$OPID[ovin_over_ten_km$DESTHOME == 0
      & ovin_over_ten_km$ORIGINHOME == 0],
82                               by = list(ovin_over_ten_km$OPID[ovin_over_ten_km$
      DESTHOME == 0 & ovin_over_ten_km$ORIGINHOME == 0]), length)
83 ovin_over_ten_km$TRIPS_NOT_HOME <- merge(ovin_over_ten_km, trips.not.home, by.x =
      "OPID", by.y = "Group.1", all = T)$x
84 # Code needed to generate this attribute at the level of the respondent
85 trips.not.home.base <- as.data.frame(last.trip.end$Group.1)
86 trips.not.home <- merge(trips.not.home, trips.not.home.base, by.x = "Group.1",
      by.y = "last.trip.end$Group.1", all = T)
87 rm(trips.not.home.base)
88 trips.not.home[is.na(trips.not.home$x),]$x <- 0
89 # Respondent attribute: Average activity duration
90 avg.activity.duration <- aggregate(ovin_over_ten_km$ACTDUUR, by = list(ovin_over_
      ten_km$OPID), mean)
91 ovin_over_ten_km$AVG_ACTIVITY_DURATION <- merge(ovin_over_ten_km,
      avg.activity.duration, by.x = "OPID", by.y = "Group.1", all = T)$x
92 # Code needed to generate this attribute at the level of the respondent
93 sd.activity.duration <- aggregate(ovin_over_ten_km$ACTDUUR, by = list(ovin_over_
      ten_km$OPID), sd)
94 ovin_over_ten_km$SD_ACTIVITY_DURATION <- merge(ovin_over_ten_km,
      sd.activity.duration, by.x = "OPID", by.y = "Group.1", all = T)$x
95 sd.activity.duration$x[is.na(sd.activity.duration$x)] <- 0
96 # Respondent attribute: Maximum activity duration

```



```

97 max.activity.duration <- aggregate(ovin_over_ten_km$ACTDUUR, by = list(ovin_over_
    ten_km$OPID), max)
98 ovin_over_ten_km$MAX_ACTIVITY_DURATION <- merge(ovin_over_ten_km,
    max.activity.duration, by.x = "OPID", by.y = "Group.1", all = T)$x
99 # Code needed to generate this attribute at the level of the respondent
100 min.activity.duration <- aggregate(ovin_over_ten_km$ACTDUUR, by = list(ovin_over_
    ten_km$OPID), min)
101 ovin_over_ten_km$MIN_ACTIVITY_DURATION <- merge(ovin_over_ten_km,
    min.activity.duration, by.x = "OPID", by.y = "Group.1", all = T)$x
102 # Respondent attribute: Number of business trips
103 sum.kmotiefv.zakelijk <- aggregate(ovin_over_ten_km$KMOTIEFV == "Zakelijk bezoek
    in werksfeer", by = list(ovin_over_ten_km$OPID), sum)
104 sum.kmotiefv.zakelijk$x[sum.kmotiefv.zakelijk$x > 0] <- 1
105 ovin_over_ten_km$SUM_KMOTIEF_ZAKELIJK <- merge(ovin_over_ten_km,
    sum.kmotiefv.zakelijk, by.x = "OPID", by.y = "Group.1", all = T)$x
106 # Respondent attribute: Number of commutes
107 sum.kmotiefv.werk <- aggregate(ovin_over_ten_km$KMOTIEFV == "Van en naar het werk
    ", by = list(ovin_over_ten_km$OPID), sum)
108 sum.kmotiefv.werk$x[sum.kmotiefv.werk$x > 0] <- 1
109 ovin_over_ten_km$SUM_KMOTIEF_WERK <- merge(ovin_over_ten_km, sum.kmotiefv.werk,
    by.x = "OPID", by.y = "Group.1", all = T)$x
110
111 # Generate ovin at the level of a respondent
112 ovin_persoon_niveau <- cbind(sum.kmotiefv.zakelijk, sum.kmotiefv.werk$x,
    first.trip.start$x, first.trip.end$x, last.trip.start$x, last.trip.end$x,
    total.trips$x, total.trips.day$x, weekend$x, trips.not.home$x,
    avg.activity.duration$x, sd.activity.duration$x, max.activity.duration$x,
    min.activity.duration$x)
113 names(ovin_persoon_niveau) <- c("OPID", "ZAKELIJK", "WERK", "FIRST_TRIP_START", "
    FIRST_TRIP_END", "LAST_TRIP_START", "LAST_TRIP_END", "TOTAL_TRIPS", "TOTAL_
    TRIPS_DAY", "WEEKEND", "TRIPS_NOT_HOME", "AVG_ACTIVITY_DURATION", "SD_ACTIVITY
    _DURATION", "MAX_ACTIVITY_DURATION", "MIN_ACTIVITY_DURATION")
114 # Clean the workspace of temporary variables
115 rm (sum.kmotiefv.zakelijk)
116 rm (sum.kmotiefv.werk)
117 rm (first.trip.start)
118 rm (first.trip.end)
119 rm (last.trip.start)

```

```
120 rm (last.trip.end)
121 rm (total.trips)
122 rm (total.trips.day)
123 rm (weekend)
124 rm (trips.not.home)
125 rm (avg.activity.duration)
126 rm (sd.activity.duration)
127 rm (max.activity.duration)
128 rm (min.activity.duration)
```

OvinCharacteristics.R

Appendix D

Factors influencing City Attractiveness

<p>Natural Resources Weather Temperature Rainfall Humidity Hours of sunshine Beaches Quality of seawater Sandy or rocky beaches Length of the beaches Overcrowding of beaches Wealth of countryside Protected nature reserves Lakes, mountains, deserts, etc. Variety and uniqueness of flora and fauna</p>	<p>General Infrastructure Development and quality of roads, airports and ports Private and public transport facilities Development of health services Development of telecommunications Development of commercial infrastructures Extent of building development</p>	<p>Tourist Infrastructure Hotel and self-catering accommodation Number of beds Categories Quality Restaurants Number Categories Quality Bars, discotheques and clubs Ease of access to destination Excursions at the destination Tourist centers Network of tourist information</p>
<p>Tourist Leisure and Recreation Theme parks Entertainment and sports activities Golf, fishing, hunting, skiing, scuba diving, etc. Water parks Zoos Trekking Adventure activities Casinos Night life Shopping</p>	<p>Culture, History and Art Museums, historical buildings, monuments, etc. Festival, concerts, etc. Handicraft Gastronomy Folklore Religion Customs and ways of life</p>	<p>Political and Economic Factors Political stability Political tendencies Economic development Safety Crime rate Terrorist attacks Prices</p>
<p>Natural Environment Beauty of the scenery Attractiveness of the cities and towns Cleanliness Overcrowding Air and noise pollution Traffic congestion</p>	<p>Social Environment Hospitality and friendliness of the local residents Underprivilege and poverty Quality of life Language barriers</p>	<p>Atmosphere of the Place Luxurious Fashionable Place with a good reputation Family-oriented destination Exotic Mystic Relaxing Stressful Fun, enjoyable Pleasant Boring Attractive or interesting</p>

TABLE D.1: Factors that influence the perceived destination image (Beerli & Martin, 2004)

Appendix E

Urban Competitiveness Model

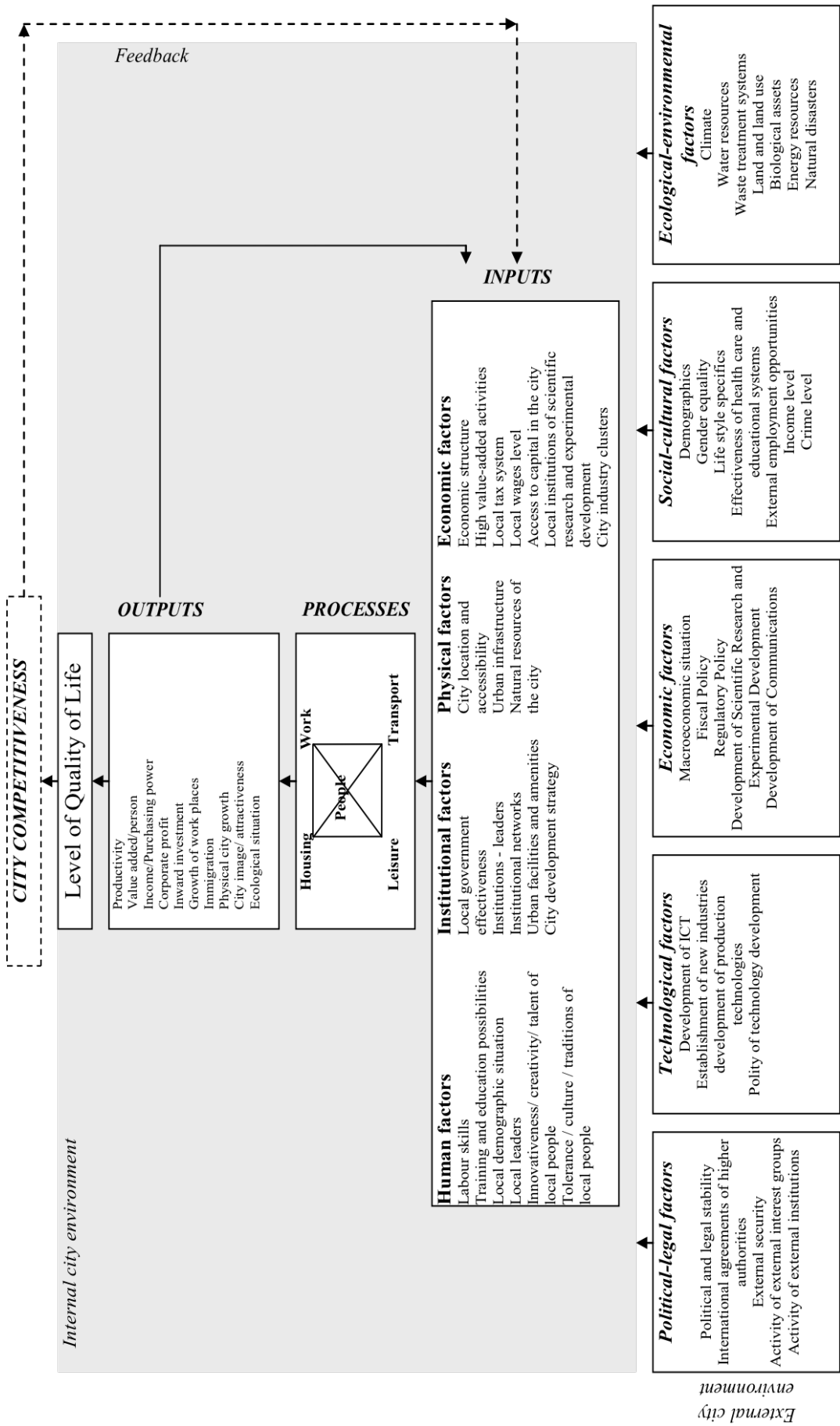


FIGURE E.1: Urban competitiveness model (Sinkienė, 2008)

Appendix F

Cell Radius Analysis

In this appendix we provide an analysis of the data quality by comparing the mobile phone location data with a [GPS](#) trace. The [GPS](#) trace covers the entire month of February 2015. The analysis will thus cover 28 days of measurement. The [GPS](#) trace is from one of the employees of Mezero for which the privacy regulations are lifted to do this and similar evaluation studies. In total four algorithms are compared against the [GPS](#) trace: the golden standard, the current algorithm, the current algorithm with only cells under 10 km, and the current algorithm with only cells under 12.5 km. These algorithms all provide us with information about the origins and destinations of the person of interest. We will evaluate the algorithms based on the number of trips acknowledged, the percentage of correctly predicted origins and the percentage of correctly predicted destinations. For the [GPS](#) trace we define a destination similar to the definition of a destination in the mobile phone location data, i.e. a person needs to be near stationary for half an hour or more. Near stationary we define as moving less than 5 km in radius. Locations are compared based on the areas also in the mobile phone location data. When a person has a destination in the [GPS](#) trace the average of the longitude and latitude at a location are mapped to these areas. We observed a total of 52 trips.

In figure [F.1](#) the percentage of trips observed for all four algorithms is depicted.

With the percentage of trips observed, i.e. the y-axis in figure [F.1](#), we mean the number of trips in the [GPS](#) data that can be traced back to the origins and destinations in the mobile phone

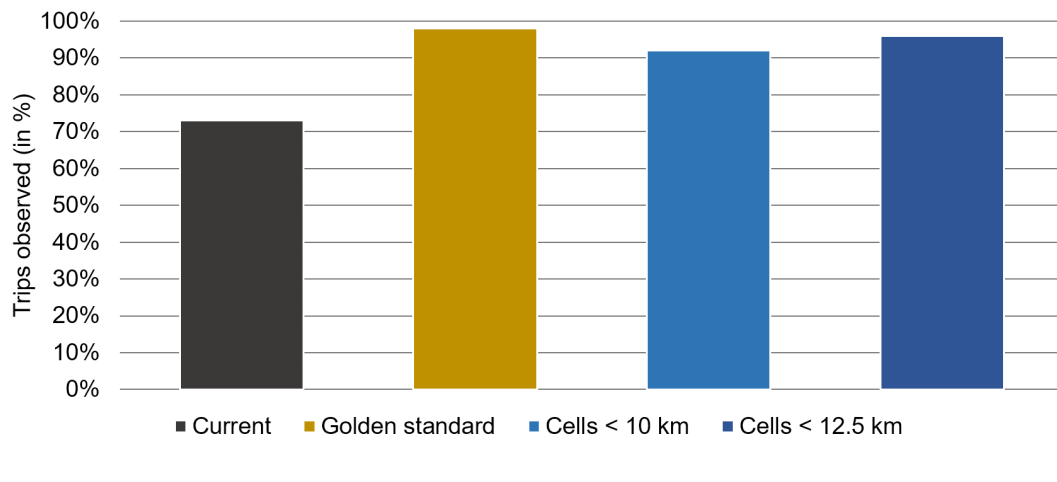


FIGURE F.1: The percentage of trips observed in the [GPS](#) trace that can also be found in the mobile phone location data.

location data. When origins and or destinations are not exactly correct, but rather positioned in neighbouring areas they are still counted.

From figure [F.1](#) we observe that the current algorithm performs much worse than the other three, with the golden standard righteously performing best. The current algorithm with only cells under 12.5 km performs nearly as good as the golden standard. The algorithm with only cells under 10 km performs slightly worse. We expect that this worse performance by the 10 km cells is caused by too many events being left out. We must note that mistakes are most often made on trips with shorter distances. Long distance trips were well recorded by all algorithms.

For all trips that can be traced back from the [GPS](#) to the mobile phone location data we also evaluated how often the origin was exactly correct, i.e. the same location as in the [GPS](#) trace (see figure [F.2](#)). We found all algorithms performed well. With the golden standard and cells under 12.5 km providing the best results. When the algorithms were incorrect, they were nearly always located in a neighbouring area. The graph with the destinations is omitted as very similar results are found.

To conclude, the current algorithm performs worse than the alternatives. The golden standard, as we expect, outperforms all algorithms. However, the idea of leaving out events with cells above a certain radius appears yield good results. Both algorithms presented in blue perform well. In particular, the current algorithm with only cells under 12.5 km appears to be nearly as good as the golden standard. It manages to observe 96% of the trips found in the [GPS](#)

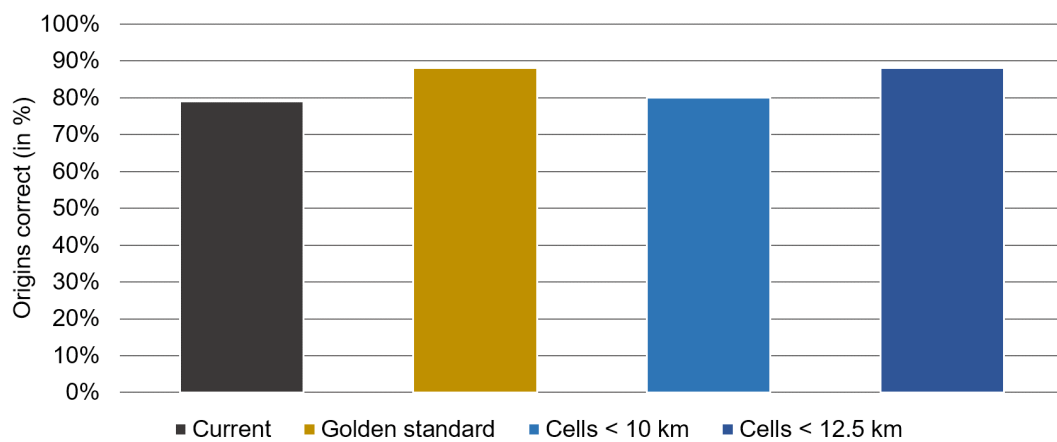


FIGURE F.2: Percentages of origins correct for all four algorithms.

compared to the 98% of the golden standard and just as often measures the correct origin for the observed trips.

Appendix G

Developing the Improved Scaling Factor

The current method used by Mezuro to extrapolate the sample to the travelling population does not take into account (1) the penetration of mobile phones per age group and (2) the likelihood that a person makes a trip greater than 10 kilometres. These distributions however, differ per age group. Moreover, the likelihood that person makes a trip greater than 10 kilometres varies. Therefore, we propose a scaling method that takes the aforementioned limitations into account.

The resulting scaling method is depicted in figure [G.1](#). Descriptions of the activities, depicted on the left hand side of this figure are provided in table [G.1](#). Detailed descriptions of the concepts, depicted on the right hand side of this figure are provided in table [G.2](#).

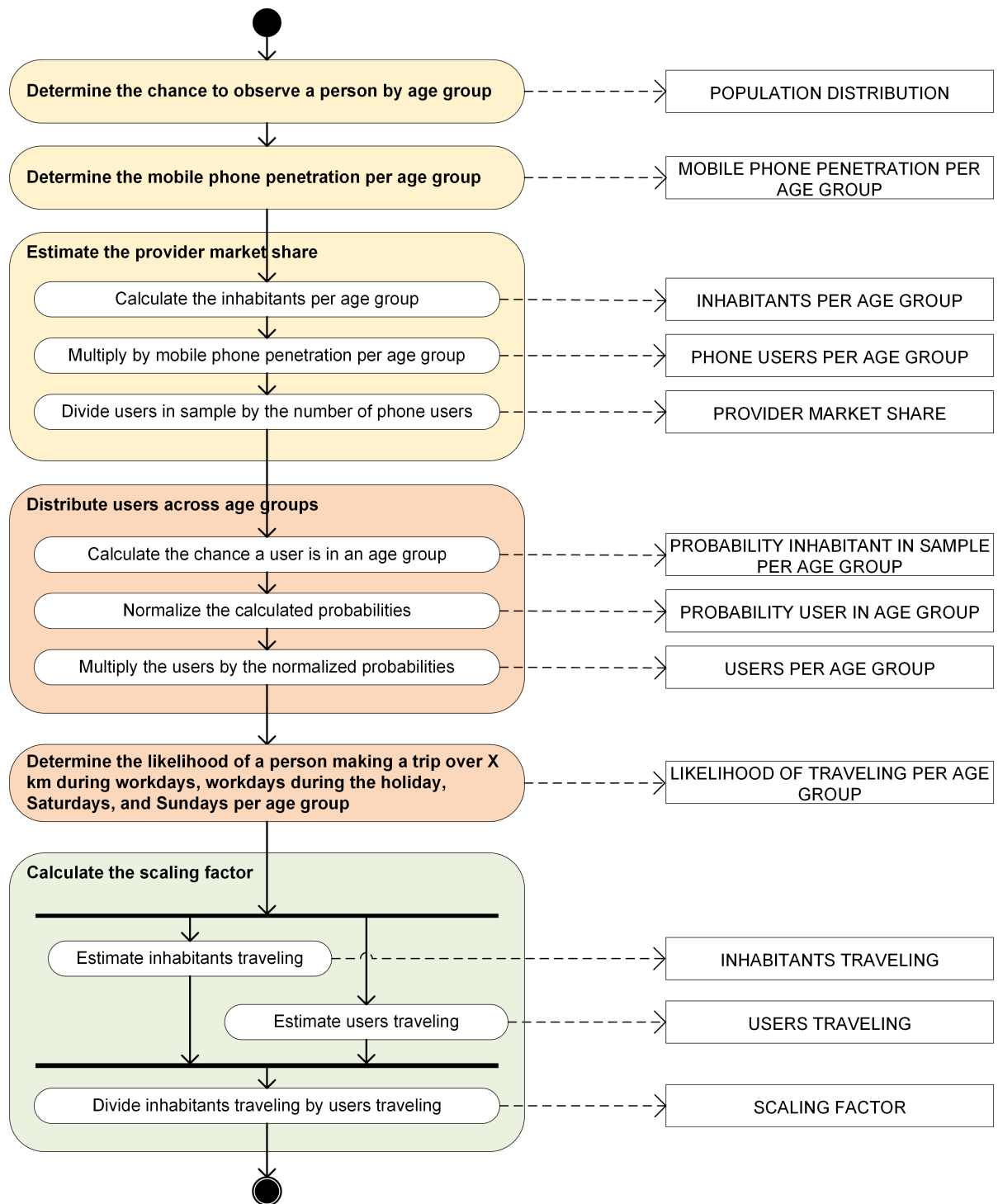


FIGURE G.1: The improved method to determine the scaling factor, which can be used to extrapolate the sample to the population

TABLE G.1: Descriptions of the activities and sub-activities that are used in the method depicted in figure G.1

Activity	Sub-activity	Description
Determine the chance to observe a person by age group		Divide the number of people per age group by the total number of inhabitants in that area.
Determine the mobile phone penetration per age group		Retrieve data concerning the mobile phone penetration per age group from (a) trusted source(s).
Estimate the provider market share	Calculate the inhabitants per age group	Multiply the total number of inhabitants by the POPULATION DISTRIBUTION. Note, the total number of inhabitants used here is adjusted for people being abroad, on a holiday or on a business trip (Geerts, 2014).
	Multiply by mobile phone penetration per age group	Multiply the INHABITANTS PER AGE GROUP by the MOBILE PHONE PENETRATION PER AGE GROUP.
	Divide users in sample by the number of phone users	Divide the number of users in the area by the number of phone users in the area, i.e. the sum of the PHONE USERS PER AGE GROUP.
Distribute users across age groups	Calculate the chance a user is in an age group	Multiply the POPULATION DISTRIBUTION by the MOBILE PHONE PENETRATION PER AGE GROUP and by the PROVIDER MARKET SHARE.
	Normalise the calculated probabilities	Divide the PROBABILITY INHABITANT IN SAMPLE PER AGE GROUP by the sum of the PROBABILITY INHABITANT IN SAMPLE PER AGE GROUP.
	Multiply the users by the normalised probabilities	Multiply the users in the area by the chance of observing a user in a certain age group, i.e. the PROBABILITY USER IN AGE GROUP.
Determine the likelihood of a person making a trip over X km during workday, workdays during the holiday, Saturdays and Sundays per age group		Gather information about the chance that a person of a certain age groups makes a trip longer than X kilometres on a day. We use OViN to determine this and take the differences in workday, workdays during the holiday, Saturdays and Sundays separately into account.
Calculate the scaling factor	Estimate inhabitants travelling	Multiply the INHABITANTS PER AGE GROUP by the LIKELIHOOD OF TRAVELLING PER AGE GROUP.
	Estimate the users travelling	Multiply the USERS PER AGE GROUP by the LIKELIHOOD OF TRAVELLING PER AGE GROUP.

Continued on the next page

TABLE G.1: (continued)

Activity	Sub-activity	Description
	Divide inhabitants travelling by users travelling	Divide the sum of the INHABITANTS TRAVELLING by the sum of the USERS TRAVELLING.

TABLE G.2: Descriptions of the concepts that are used in the method depicted in figure G.1

Concept	Description
POPULATION DISTRIBUTION	The probability that an inhabitant belongs to a certain age group.
MOBILE PHONE PENETRATION PER AGE GROUP	The probability that a Dutch citizen of a certain age group possesses a mobile phone.
INHABITANTS PER AGE GROUP	The absolute number of inhabitants per age group.
PHONE USERS PER AGE GROUP	The absolute number of inhabitants that possess a mobile phone per age group.
PROVIDER MARKET SHARE	The market share of the network provider. Hence, in this case the market share is equal for all age groups.
PROBABILITY INHABITANT IN SAMPLE PER AGE GROUP	The probability that an inhabitant is in our sample per age group.
PROBABILITY USER IN AGE GROUP	The probability that a user in our sample is in a certain age group.
USERS PER AGE GROUP	The absolute number of users per age group.
LIKELIHOOD OF TRAVELLING PER AGE GROUP	The probabilities that a person of a certain age group makes a trip that is longer than X kilometre on a specific day of the week.
INHABITANTS TRAVELLING	The number of INHABITANTS PER AGE GROUP that is expected to make a trip over X kilometres on a specific day of the week.
USERS TRAVELLING	The number of USERS PER AGE GROUP that is expected to make a trip over X kilometres on a specific day of the week.
SCALING FACTOR	The ratio between the number of INHABITANTS TRAVELLING the number of USERS TRAVELLING. The scaling factors applies to travelling people, because these are the people that are relevant for the OD matrix.

Appendix H

Regression Results

This appendix shows the results of the regression models per trip motive. The tables [H.1](#), [H.2](#) and [H.3](#) use the A_j , P_j and R_j measure of city attractiveness as dependent variable, respectively, for each trip motive. The measures of city attractiveness are explained in detail in section [7.4](#). The relative importance derived from these models is presented in appendix [K](#).

TABLE H.1: Regression results per trip motive for the absolute measure city attractiveness (A_j)

	<i>Dependent variable:</i>								
	Home to work (1)	Business (2)	Services (3)	Shopping (4)	Education (5)	Visiting (6)	Social recr. other (7)	Touring (8)	Other (9)
City size	657,066.800*** (56,273.980)	123,215.500*** (10,665.540)	58,460.900*** (5,099.732)	134,205.500*** (11,708.830)	110,139.100*** (9,778.060)	244,571.500*** (22,512.830)	237,333.800*** (21,832.060)	25,140.600*** (2,330.567)	141,766.300*** (12,172.260)
City type	-121,078.700**	-17,166.910	-9,492.383*	-13,939.360	-20,480.030**	-37,636.780	-35,453.600	-3,959.386	-17,405.100
Wealth	64,090.550	16,225.240	3,813.818	9,388.796	10,573.450	23,312.120	20,433.280	2,486.103	15,851.890
Elderly	-112,780.300*	-31,630.060***	-11,764.020**	-31,489.340**	-19,521.450*	-53,763.070**	-50,595.160**	-5,246.985**	-35,107.700***
Single HHs	73,064.870	5,549.701	4,968.026	4,473.531	12,555.460	15,591.410	16,324.820	1,588.790	6,445.067
Constant	559,121.000*** (55,328.130)	142,036.800*** (10,486.270)	55,956.560*** (5,014.016)	147,029.700*** (11,512.020)	93,914.190*** (9,613.711)	250,931.800*** (22,134.440)	242,209.100*** (21,465.110)	24,687.870*** (2,291.395)	161,304.400*** (11,967.670)
Observations	30	30	30	30	30	30	30	30	30
R ²	0.860	0.860	0.855	0.854	0.852	0.842	0.842	0.840	0.860
Adjusted R ²	0.831	0.831	0.825	0.824	0.821	0.809	0.809	0.807	0.831
F-Statistic (df = 5; 24)	29.593***	29.487***	28.341***	28.163***	27.613***	25.614***	25.524***	25.185***	29.597***

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE H.2: Regression results per trip motive for the measure city attractiveness per inhabitant (P_j)

	<i>Dependent variable:</i>								
	Home to work (1)	Business (2)	Services (3)	Shopping (4)	Education (5)	Visiting (6)	Social recr. other (7)	Touring (8)	Other (9)
City size	0.689** (0.272)	0.044 (0.046)	0.047* (0.023)	0.055 (0.048)	0.118** (0.048)	0.147 (0.098)	0.149 (0.095)	0.017 (0.010)	0.051 (0.051)
City type	-0.864*** (0.272)	-0.134*** (0.046)	-0.072*** (0.023)	-0.101** (0.048)	-0.153*** (0.048)	-0.269** (0.098)	-0.257** (0.095)	-0.028** (0.010)	-0.128** (0.051)
Wealth	-0.110 (0.272)	0.010 (0.046)	-0.017 (0.023)	-0.032 (0.048)	-0.016 (0.048)	-0.031 (0.098)	-0.042 (0.095)	-0.003 (0.010)	-0.003 (0.051)
Elderly	0.259 (0.272)	0.022 (0.046)	0.017 (0.023)	0.024 (0.048)	0.040 (0.048)	0.033 (0.098)	0.044 (0.095)	0.004 (0.010)	0.024 (0.051)
Single HHs	0.769*** (0.272)	0.077 (0.046)	0.059** (0.023)	0.090* (0.048)	0.131** (0.048)	0.185* (0.098)	0.194* (0.095)	0.018* (0.010)	0.098* (0.051)
Constant	2.793*** (0.267)	0.785*** (0.045)	0.293*** (0.023)	0.811*** (0.047)	0.464*** (0.047)	1.339*** (0.097)	1.293*** (0.094)	0.130*** (0.010)	0.892*** (0.050)
Observations	30	30	30	30	30	30	30	30	30
R ²	0.516	0.340	0.465	0.291	0.506	0.359	0.372	0.356	0.318
Adjusted R ²	0.416	0.202	0.353	0.144	0.404	0.226	0.241	0.221	0.176
F Statistic (df = 5, 24)	5.124***	2.470*	4.169***	1.974	4.925***	2.691**	2.844**	2.648**	2.240*

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE H.3: Regression results per trip motive for the ratio between attracted and non-retained visits (R_j)

	<i>Dependent variable:</i>								
	Home to work (1)	Business (2)	Services (3)	Shopping (4)	Education (5)	Visiting (6)	Social recr. other (7)	Touring (8)	Other (9)
City size	0.340*** (0.109)	0.439* (0.213)	0.221** (0.106)	0.208 (0.128)	0.285** (0.106)	0.205** (0.092)	0.211** (0.094)	0.219** (0.090)	0.293* (0.150)
City type	-0.361*** (0.109)	-0.649*** (0.213)	-0.356*** (0.106)	-0.386*** (0.128)	-0.347*** (0.106)	-0.251** (0.092)	-0.267*** (0.094)	-0.242** (0.090)	-0.423*** (0.150)
Wealth	-0.215* (0.109)	-0.486** (0.213)	-0.238** (0.106)	-0.295** (0.128)	-0.198* (0.106)	-0.155 (0.092)	-0.170* (0.094)	-0.139 (0.090)	-0.303* (0.150)
Elderly	0.145 (0.109)	0.354 (0.213)	0.135 (0.106)	0.195 (0.128)	0.133 (0.106)	0.125 (0.092)	0.128 (0.094)	0.112 (0.090)	0.258* (0.150)
Single HHs	0.337*** (0.109)	0.540** (0.213)	0.310*** (0.106)	0.409*** (0.128)	0.300*** (0.106)	0.273*** (0.092)	0.292*** (0.094)	0.248** (0.090)	0.477*** (0.150)
Constant	2.146*** (0.108)	4.811*** (0.210)	2.303*** (0.104)	3.129*** (0.126)	2.048*** (0.104)	2.377*** (0.090)	2.369*** (0.092)	2.246*** (0.089)	3.727*** (0.148)
Observations	30	30	30	30	30	30	30	30	30
R ²	0.598	0.537	0.563	0.552	0.563	0.521	0.538	0.506	0.546
Adjusted R ²	0.514	0.441	0.472	0.459	0.473	0.422	0.441	0.403	0.451
F Statistic (df = 5, 24)	7.127***	5.574***	6.190***	5.913***	6.196***	5.226***	5.579***	4.916***	5.771***

Note: *p<0.1; **p<0.05; ***p<0.01

Appendix I

Attribute and Data Source Mapping

The goal is to systematically, transparently and comprehensively assess all factors that can influence the destination image. We do this by adopting the list of factors that can influence the destination image formation from [Beerli and Martin \(2004\)](#) and mapping representative data to each factor.

In the first column of table [I.2](#) are the factors we've adopted from [Beerli and Martin \(2004\)](#). The bold terms in the first, i.e. factor, column indicate the category of the related factors. The second column shows the relevance of the corresponding factor expressed as a -1, 0 or 1. This indicates whether a factor is considered irrelevant, relevant but no representative information was found, or relevant and representative information has been found, respectively. If the factor column contains a 1, the source column shows which source the data has been obtained from. The number in the source column can be cross-referenced with table [I.1](#). The fourth column shows the column name that the information was obtained from (in the data source). The fifth column shows the unit of measurement as to how it is included in our analysis. The last column provides an explanation for the factors that are excluded from the analysis, i.e. where the relevance column is equal to -1.

TABLE I.1: These data sources and tables are used in mapping the factors that influence the attractiveness of a destination to data to represent those factors.

ID	Dataset	Source	Year	Table name
1	neighbourhood	CBS	2013	Kerncijfers wijken en buurten
2	land use	CBS	2010	Bodemgebruik; wijk- en buurtcijfers
3	amenities	CBS	2013	Nabijheid voorzieningen; afstand locatie, wijk- en buurtcijfers
4	religion	CBS	2013	Kerkelijke gezindte en kerkbezoek
5	registered criminality	CBS	2013	Geregistreeerde criminaliteit; soort misdrijf en regio
6	mobile phone location data	Mezuro	2015	October 2015
7	parking fees	PP.nl	2015	PrettigParkeren.nl

TABLE I.2: This comprehensive list of factors from Beerli and Martin (2004) that influence destination attractiveness shows how data is mapped to each factor. For each factor the relevance of this factor for this research, the source and column the data is obtained from, how the factor is implemented and expressed in this research, is elaborated.

Factor	Relevance	Source	Column	Implementation	Explanation
Natural sources	Re-				
Weather	-1				Irrelevant, too little variation when aggregated over a month
Temperature	-1				Irrelevant, too little variation when aggregated over a month
Rainfall	-1				Irrelevant, too little variation when aggregated over a month
Humidity	-1				Irrelevant, too little variation when aggregated over a month
Hours of sunshine	-1				Irrelevant, too little variation when aggregated over a month
Beaches	1	2	Bos en open natuurlijk terrein >Open droog natuurlijk terrein	ha of beach, dunes and heather	
Quality of seawater	-1				Irrelevant, too little variation within the Netherlands
Sandy or rocky beaches	-1				Irrelevant, too little variation within the Netherlands
Length of the beaches	1	2	Bos en open natuurlijk terrein >Open droog natuurlijk terrein	ha of beach, dunes and heather	
Overcrowding of beaches	0				Not implemented due to a lack of data
Wealth of countryside	1	2	Agrarisch terrein >Overig agrarisch terrein	ha of rural area	
Protected nature reserves	1	2	Bos en open natuurlijk terrein >Bos	ha of wooded area	
Lakes, mountains, deserts, etc.	1	2	Binnenwater >Totaal binnenwater	ha of inland waters	
Variety and uniqueness of flora and fauna	0				Not implemented due to a lack of data
General Infrastructure					

Continued on the next page

TABLE I.2: (continued)

Variable	Relevance	Source	Column	Implementation	Explanation
Development and quality of roads, airports and ports	1	2	Verkeersterrein, Verkeer en Vervoer >afstand tot oprit hoofdverkeersweg	ha of infrastructure & distance to nearest on-ramp	
Private and public transport facilities	1	3	Verkeer en vervoer >treinstations	average distance to nearest train station	
Development of health services	1	1	Nabijheid voorzieningen >afstand tot huisartsenpraktijk	average distance to the nearest primary care facility	
Development of telecommunications	-1				Irrelevant, national GSM network coverage is 99,9%
Development of commercial infrastructures	1	2	Verkeersterrein >wegverkeersterrein	ha of land used for infrastructure	
Extent of building development	1	2	Verkeersterrein >wegverkeersterrein	ha of land used for infrastructure	
Tourist Infrastructure					
Hotel and self-catering accommodation	1	3	horeca >hotels en dergelijken	number hotels within the proximity and distance to closest hotel	
Number of beds	0				Not implemented due to a lack of data
Categories	0				Not implemented due to a lack of data
Quality	0				Not implemented due to a lack of data
Restaurants	1	3	horeca >restaurants	number of restaurants within 1km	
Number	1	3	horeca >restaurants	number of restaurants within 1km	
Categories	0				Not implemented due to a lack of data
Quality	0				Not implemented due to a lack of data
Bars, discotheques and clubs	1	3	cafes	distance to cafes	
Ease of access to destination	1	2	verkeersterrein >wegverkeersterrein	ha of land used for infrastructure	
Excursions at the destination	0				Not implemented due to a lack of data
Tourist centres	0				Not implemented due to a lack of data
Network of tourist information	0				Not implemented due to a lack of data
Leisure and Recreation					
Theme parks	1	3	attractions	number of attractions within 10km	
Entertainment and sports activities	1	2	recreatieterrein >sportterrein	ha of sport terrain	
Golf, fishing, hunting, skiing, scuba diving, etc.	1	2	recreatieterrein >sportterrein	ha of sport terrain	
Water parks	1	3	attractions	number of attractions within 10km	
Zoos	1	2	recreatieterrein >dagrecreatief terrein	ha of daily recreation terrain (includes zoos)	

Continued on the next page

TABLE I.2: (continued)

Variable	Relevance	Source	Column	Implementation	Explanation
Trekking	1	2	recreatieterrein >verblijfsrecreatief terrein	ha of campings etc.	
Adventure activities	0				Not implemented due to a lack of data
Casinos	0				Not implemented due to a lack of data
Night life	0				Not implemented due to a lack of data
Shopping	1	2	bebouwd terrein >terrein voor detailhandel en horeca	ha of retail	
Culture, History and Art					
Museums, historical buildings, monuments, etc.	1	3	number of museums within 5km, number of cinemas within 5km	number of museums within 5km, number of cinemas within 5km	
Festival, concerts, etc.	1	3	number of podiums within 5km, number of poppodiums within 5km	number of podiums within 5km, number of poppodiums within 5km	
Handicraft	1	2	bebouwd terrein >terrein voor sociaal culturele voorzieningen, cafeteria, cafes	ha of social cultural area	
Gastronomy	1	3		number of cafes within 1km, number of cafeterias within 1km,	
Folklore	0				Not implemented due to a lack of data
Religion	1	4	religious visits and percentage of the population religious	average frequency of religious visits, percentage of population that is religious	
Customs and ways of life	0				Irrelevant, too little variation within the Netherlands
Political and Economic Factors					
Political stability	-1				These factors are mostly relevant at an international scale
Political tendencies	-1				Irrelevant, too little variation within the Netherlands
Economic development	1	1	business establishments	business establishments per line of business	
Safety	0				Data is only available on province level
Crime rate	1	5	registered crime	number of registered crimes per 1000 inhabitants	
Terrorist attacks	-1				Irrelevant, too little variation within the Netherlands
Prices	-1				Irrelevant, too little variation within the Netherlands
Natural Environment					
Beauty of the scenery	1	2	recreatieterrein >park en plantsoen		

Continued on the next page

TABLE I.2: (continued)

Variable	Relevance	Source	Column	Implementation	Explanation
Attractiveness of the cities and towns	1	6	city attractiveness	calculated column	
Cleanliness	0				Not implemented due to a lack of data
Overcrowding	0				Not implemented due to a lack of data
Air and noise pollution	0				Not implemented due to a lack of data
Traffic congestion	0				Not implemented due to a lack of data
Social Environment	-1				Irrelevant, too little variation within the Netherlands
Hospitality and friendliness of the local residents	0				Not implemented due to a lack of data
Underprivilege and poverty	-1				Irrelevant, too little variation within the Netherlands
Quality of life	0				A very complex concept, which is hard to measure
Language barriers	-1				Irrelevant, too little variation within the Netherlands
Atmosphere of the Place					
Luxurious	0				Not implemented due to a lack of data
Fashionable	0				Not implemented due to a lack of data
Attractive or interesting	0				Not implemented due to a lack of data

Appendix J

Descriptive Statistics

The descriptive statistics of the variables included in the [PCA](#) analysis are shown in table [J.2](#). Between parentheses is stated how the variable is included. This can be as a percentage of the number of inhabitants (% of Pop.), a percentage of the number of businesses (% of B.), a percentage of the number of households (% of HH.), a percentage of the number of houses (% of H.), a percentage of the total area (% of total area), the average number of amenities of that type within a radius of X kilometres of every address in the city (within X km), as the average distance in kilometres from every address in the city to the nearest amenity of that type (distance to in km), or as the average number of amenities of that type between a radius of X1 and X2 kilometres of every address in the city (X1km to X2km).

The colour codes presented in table [J.1](#) are used to cross-reference the variables presented in table [J.2](#) to the literature sources that were identified from the literature review in chapter [4](#).

TABLE J.1: Colors to identify the literature source.





Color	Literature source
	Determinants that influence the destination attractiveness (Beerli & Martin, 2004)
	Socio-demographic and socio-economic factors (Martin et al., 1998)
	Competitive forces (Deas & Giordano, 2001)
	Determinants of city centre attractiveness (Teller & Reutterer, 2008)

TABLE J.2: Descriptive statistics for the variables that characterise a city.

Variable	N	Mean	Std.Dev.	Median	Min	Max
Inhabitants	30	178253.00	154341.30	128055.00	37400.00	715550.00
Men (% of Pop.)	30	49.31	0.96	49.12	47.85	53.12
Women (% of Pop.)	30	50.68	0.96	50.86	46.87	52.14
Birth rate (% of Pop.)	30	1.10	0.17	1.08	0.74	1.54
Death rate (% of Pop.)	30	0.82	0.15	0.81	0.48	1.19
Households (% of Pop.)	30	49.10	4.32	48.75	41.31	59.82
Houses (% of Pop.)	30	46.33	2.65	46.29	39.76	52.03
Income earners (% of Pop.)	30	70.46	1.83	70.82	66.23	74.29
Unemployment rate (% of Pop.)	30	2.52	0.36	2.46	1.69	3.22
Old-age pension rate (% of Pop.)	30	14.97	2.72	14.63	8.62	20.96
Business establishments	30	14886.83	17776.09	9590.00	2635.00	93385.00
A Agriculture, forestry and fishing (% of B.)	30	0.80	0.86	0.49	0.00	3.69
B-F Industry and energy (% of B.)	30	12.87	2.78	12.80	7.25	21.19
G+I Wholesale and retail trade (% of B.)	30	23.96	3.70	23.45	17.57	32.35
H+J Transport, information & communication	30	9.65	1.79	9.48	6.58	13.47
K-L Finance and real estate (% of B.)	30	9.37	1.42	9.14	7.28	12.53
M-N Business services (% of B.)	30	28.91	3.78	29.27	20.10	35.80
R-U Culture, recreation and other services	30	14.32	2.04	13.80	11.19	19.44
Cars (% of Pop.)	30	40.97	5.87	41.46	27.58	50.39
0-15 years (% of Pop.)	30	16.32	2.13	16.40	11.56	20.64
15-25 years (% of Pop.)	30	13.85	3.12	12.69	10.56	23.88
25-45 years (% of Pop.)	30	28.80	3.02	28.97	23.47	36.61
45-65 years (% of Pop.)	30	25.99	2.35	25.68	19.82	31.09
65 years and older (% of Pop.)	30	15.05	2.77	14.60	8.37	21.17
Single (% of Pop.)	30	52.02	5.70	51.32	42.64	66.00
Married (% of Pop.)	30	34.97	4.71	35.62	23.35	41.68
Divorced (% of Pop.)	30	8.20	0.93	8.36	6.01	9.86
Widowed (% of Pop.)	30	4.75	1.01	4.84	2.86	7.40
Single households (% of HH.)	30	42.77	7.46	42.88	30.30	57.99
Households without kids (% of HH.)	30	26.01	2.83	26.33	20.63	31.44
Households with kids (% of HH.)	30	31.18	5.63	30.95	20.35	45.19
Average household size	30	2.06	0.17	2.05	1.72	2.39
Average house value (x1,000)	30	198.62	31.18	191.66	141.51	269.99
Single-family housing (% of H.)	30	56.37	16.53	59.50	12.00	85.46
Multi-family housing (% of H.)	30	43.70	16.45	40.50	15.23	88.00
Occupied housing (% of H.)	30	94.98	1.53	95.30	90.68	97.16
Vacant housing (% of H.)	30	5.02	1.53	4.70	2.84	9.32
Private housing (% of H.)	30	48.39	8.59	48.81	25.48	63.96
Rented housing (% of H.)	30	50.62	8.16	50.44	35.84	73.40
Avg. electricity usage (kWh)	30	2929.69	280.07	2982.89	2309.27	3440.15
Avg. natural gas usage (m ³)	30	1414.51	277.27	1490.46	557.02	1909.97

Continued on the next page

TABLE J.2: (continued)

Variable	N	Mean	Std.Dev.	Median	Min	Max
Avg. income per income earner (x1,000 euro)	30	30.30	2.52	30.37	26.22	35.58
Avg. income per resident (x1,000 euro)	30	21.88	1.73	21.93	18.85	25.67
Low income residents (% of Pop.)	30	40.69	4.00	40.10	33.34	47.80
High income residents (% of Pop.)	30	19.35	3.45	19.54	13.65	25.31
Inactive residents (% of Pop.)	30	25.47	3.92	24.97	18.86	35.14
Low income households (% of HH.)	30	44.74	5.26	44.99	33.95	54.49
High income households (% of HH.)	30	17.50	3.48	17.32	12.30	25.22
Urbanity	30	2494.70	1059.41	2215.45	969.00	6156.66
Crime rate (per 1,000 Pop.)	30	84.17	16.61	83.10	52.80	124.40
Frequency of religious visits	30	11.77	3.34	11.00	7.70	19.50
Religious population (% of Pop.)	30	43.38	10.56	41.35	27.30	72.50
Parking fees (euro)	30	2.70	0.77	2.60	1.30	5.00
Traffic area (% of total area)	30	6.30	1.49	6.32	3.23	9.89
Railroad area (% of total area)	30	0.99	0.52	0.93	0.24	2.58
Road area (% of total area)	30	5.09	1.32	5.01	2.85	8.36
Airport area (% of total area)	30	0.21	0.47	0.00	0.00	1.78
Urban area (% of total area)	30	38.30	12.98	38.88	8.06	59.06
Residential area (% of total area)	30	24.41	9.14	24.22	4.82	40.85
Retail and hospitality industry area (% of area)	30	1.57	0.79	1.43	0.27	3.32
Public facilities area (% of total area)	30	1.06	0.62	1.00	0.19	2.58
Sociocultural facilities area (% of total area)	30	2.50	1.33	2.37	0.52	5.77
Commercial area (% of total area)	30	8.76	4.64	8.49	1.67	25.58
Recreational area (% of total area)	30	9.21	3.83	8.69	3.52	17.13
Parks and greenspace area (% of total area)	30	4.65	2.44	4.35	1.29	10.09
Sports area (% of total area)	30	3.05	1.12	2.97	0.83	5.09
Rural area (% of total area)	30	24.61	13.77	23.62	0.91	56.12
Countryside area (% of total area)	30	24.38	13.76	23.11	0.91	56.11
Forest and other natural area (% of total area)	30	10.70	10.76	5.33	0.58	41.96
Forest area (% of total area)	30	8.40	8.46	4.19	0.50	32.42
Dry natural area (% of total area)	30	1.15	2.40	0.08	0.00	8.48
Wet natural area (% of total area)	30	1.14	2.37	0.34	0.00	10.45
Inland water area (% of total area)	30	6.57	5.81	5.54	0.44	25.14
Primary care facilities (within 1km)	30	2.19	1.09	1.85	0.77	5.49
Hospitals (within 5km)	30	1.55	0.88	1.33	0.43	4.34
Supermarkets (within 1km)	30	2.37	0.88	2.23	1.09	5.24
Other retail (within 1km)	30	12.27	8.41	9.79	2.75	39.27
Department stores (within 5km)	30	4.67	2.45	4.10	1.80	12.18
Cafes (within 1km)	30	8.57	6.29	8.15	0.82	29.30
Cafeterias (within 1km)	30	9.31	4.89	8.92	2.32	25.04
Restaurants (within 1km)	30	13.07	9.92	9.46	2.06	52.67
Hotels (within 5km)	30	18.32	38.92	7.78	1.34	218.53

Continued on the next page

TABLE J.2: (continued)

Variable	N	Mean	Std.Dev.	Median	Min	Max
Highway (distance to in km.)	30	1.96	0.49	1.97	1.05	3.00
Railway station (distance to in km.)	30	2.29	0.51	2.30	1.53	3.32
Libraries (distance to in km.)	30	1.59	0.54	1.43	1.02	3.02
Pools (distance to in km.)	30	2.61	0.93	2.33	1.41	5.61
Ice rinks (distance to in km.)	30	13.12	12.23	6.64	2.49	52.62
Museums (within 5km)	30	5.53	5.69	4.61	0.06	29.33
Cinemas (within 5km)	30	2.02	1.32	1.96	0.00	6.43
Amusement parks (within 10km)	30	2.37	1.34	2.27	0.00	6.08
Podium arts (within 5km)	30	3.95	4.28	2.91	0.69	23.86
Peripheral amusement parks (10km to 50km)	30	24.38	10.37	22.25	5.51	41.87
Peripheral cinemas (5km to 20km)	30	5.34	4.69	3.61	0.00	16.70
Peripheral cafes (1km to 5km)	30	91.81	98.95	65.98	11.13	504.14
Peripheral cafeteria (1km to 5km)	30	105.46	84.73	78.98	28.72	434.11
Peripheral hotels (5km to 20km)	30	79.15	90.28	39.08	2.74	403.67
Peripheral primary care facilities (1km to 5km)	30	25.90	17.57	21.67	6.42	82.97
Peripheral museums (5km to 20km)	30	23.65	18.36	16.68	0.55	72.27
Peripheral podium arts (5km to 20km)	30	12.64	13.71	7.00	0.03	56.09
Peripheral restaurants (1km to 5km)	30	149.97	169.00	103.44	32.77	928.45
Peripheral supermarkets (1km to 5km)	30	27.93	14.43	24.16	11.30	76.83
Peripheral department stores (5km to 20km)	30	21.02	15.80	15.85	1.08	58.06
Peripheral other retail (1km to 5km)	30	136.76	129.23	95.71	34.75	601.41

Appendix K

Relative importance per trip motive

The relative importance of the factors that influence city attractiveness are based on the R^2 , i.e. the explanatory value, of the predictors on the models shown in Appendix H. The results below are obtained by feeding the previously mentioned regression models to the Relaimpo package in R (Grömping et al., 2006). More specifically, the `calc.relimp()` function within this package is used to obtain the relative importance. The method to determine the relative importance is called 'first'. Most of the alternative methods to determine the relative importance that are available by the Relaimpo package are developed to deal with a certain amount of correlation between the independent variables (Grömping et al., 2006). However, since the independent variables in our models are uncorrelated we can use the most basic method to determine the relative importance of our predictors, which is 'first'. The result of the 'first' method is equal to the squared correlation coefficient between each independent (i.e., the five principal components) and dependent variable (i.e., the measure of city attractiveness).

Table K.1, K.2 and K.3 shows the relative importance of the components per trip motive for the three measures of city attractiveness A_j , P_j and R_j , respectively. These measures are explained in detail in section 7.4. A graphical representation of these three tables can be found in the results, i.e. chapter 8 figure 8.5.

TABLE K.1: Relative importance of the components per trip motive for the absolute measure city attractiveness (A_j)

	City size	City type	Wealth	Elderly	Single HHs
Other	91,7%	1,4%	1,1%	5,6%	0,2%
Services	92,7%	2,4%	0,4%	3,8%	0,7%
Education	91,9%	3,2%	0,8%	2,9%	1,2%
Social recr.	92,6%	2,1%	0,7%	4,2%	0,4%
Touring	92,4%	2,3%	0,9%	4,0%	0,4%
Work	92,1%	3,1%	0,9%	2,7%	1,1%
Visiting	92,2%	2,2%	0,8%	4,5%	0,4%
Shopping	93,3%	1,0%	0,5%	5,1%	0,1%
Business	90,5%	1,8%	1,6%	6,0%	0,2%

TABLE K.2: Relative importance of the components per trip motive for the measure city attractiveness per inhabitant (P_j)

	City size	City type	Wealth	Elderly	Single HHs
Other	8,9%	56,3%	0,0%	1,9%	32,8%
Services	19,2%	45,2%	2,4%	2,6%	30,7%
Education	24,8%	41,7%	0,5%	2,8%	30,3%
Social recr.	17,2%	51,1%	1,3%	1,5%	28,9%
Touring	20,4%	54,4%	0,7%	1,1%	23,4%
Work	25,1%	39,5%	0,6%	3,5%	31,2%
Visiting	16,6%	55,6%	0,7%	0,8%	26,2%
Shopping	13,2%	44,2%	4,6%	2,4%	35,6%
Business	7,2%	67,9%	0,4%	1,8%	22,6%

TABLE K.3: Relative importance of the components per trip motive for the ratio between attracted visits and non-retained residents (R_j)

	City size	City type	Wealth	Elderly	Single HHs
Other	13,2%	27,6%	14,1%	10,2%	34,9%
Services	14,1%	36,6%	16,3%	5,3%	27,7%
Education	23,2%	34,6%	11,3%	5,1%	25,8%
Social recr.	18,1%	29,0%	11,7%	6,6%	34,6%
Touring	24,0%	29,4%	9,7%	6,2%	30,8%
Work	27,0%	30,6%	10,9%	4,9%	26,6%
Visiting	19,2%	28,6%	11,0%	7,2%	34,0%
Shopping	8,9%	30,7%	17,9%	7,9%	34,5%
Business	15,2%	33,2%	18,6%	9,9%	23,0%

Appendix L

City Area Selection

It is important to document the area that is included per city as the results are dependent upon how the cities are selected. The area included and excluded differs per data source as the geographical boundaries do not always match. Hence, in the figures below different colours are used to illustrate which included areas belongs to which dataset. Two datasets are distinguished in the figures below: the mobile phone location data and the [CBS](#) data. The pictures consist of four layers and were made using QGIS (a graphical information system). These layers are: a map of the Netherlands, the Meuzuro areas, [CBS](#) districts, and [CBS](#) neighbourhoods. These last two layers can be seen as derived from the same data, however the neighbourhoods data is finer grained, in terms of the geographical area covered, than the districts data.

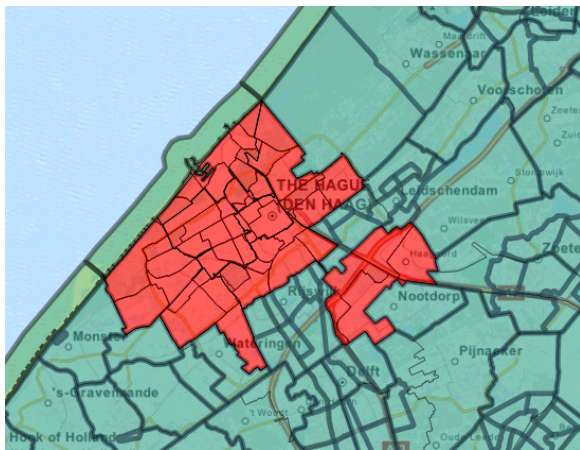
The steps that were performed to select the areas for each city included in this study are:

1. Find the city on Google Maps
2. Select the Meuzuro area(s) to cover the area of the city as specified by Google Maps
3. Select the [CBS](#) district layers to match the area covered by the Meuzuro area(s) areas
4. Select any missing areas using the [CBS](#) neighbourhoods layer

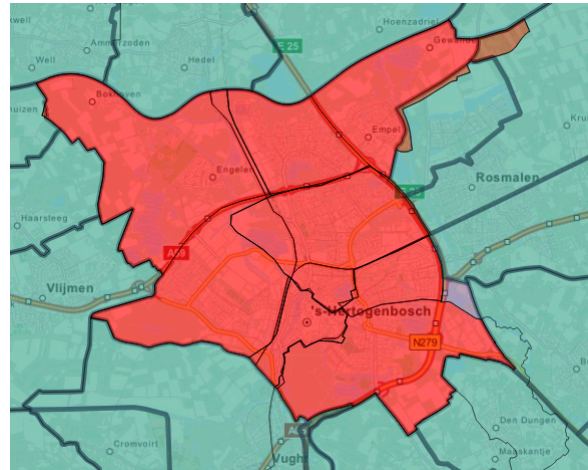
More than once it was impossible to make an identical selection using the layers. In this case the selection was considered to be a trade-off, i.e. if the Meuzuro areas covered a little bit of extra area compared to the [CBS](#) area, the [CBS](#) area was allowed to cover a little bit of area not

covered by the Mezero area. For an example see the figure of 's-Hertogenbosch below. In this figure the orange and purple area are the CBS neighbourhoods and Mezero area, respectively. This way the total surface covered per city remained largely the same across the datasets.

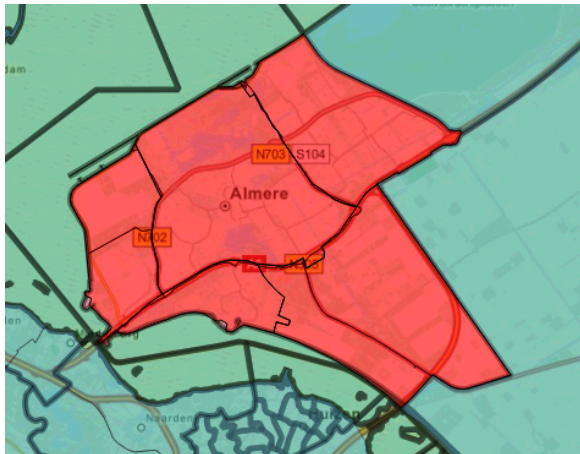
In the remainder of this appendix will be figures that graphically depict the selected area per city. In these figures the green/turquoise coloured area—the area that fully surrounds almost all of the cities—is area that does not belong to that particular city. The red coloured area is area that does belong to that particular city and is used in the CBS dataset and the mobile phone location data. The purple coloured area is area that belongs only to the mobile phone location data. The brownish coloured area is area from the CBS dataset. In summary, all except the green/turquoise coloured area belongs to that particular city.



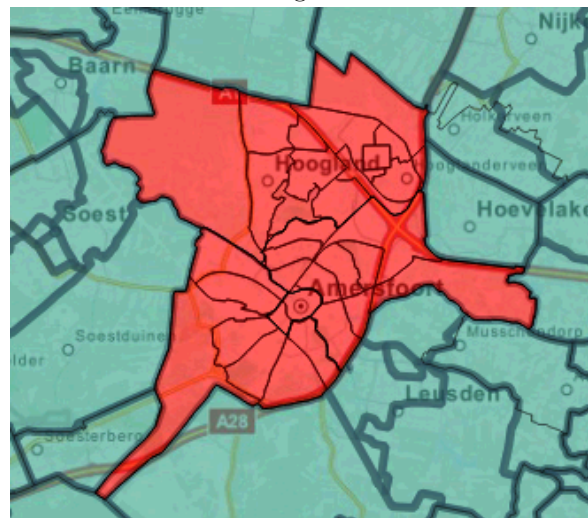
's-Gravenhage



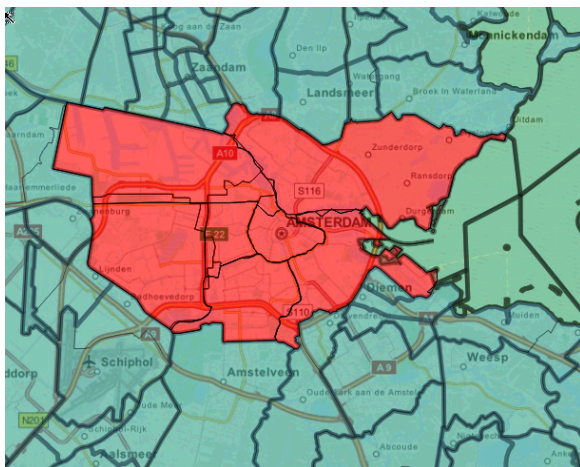
's-Hertogenbosch



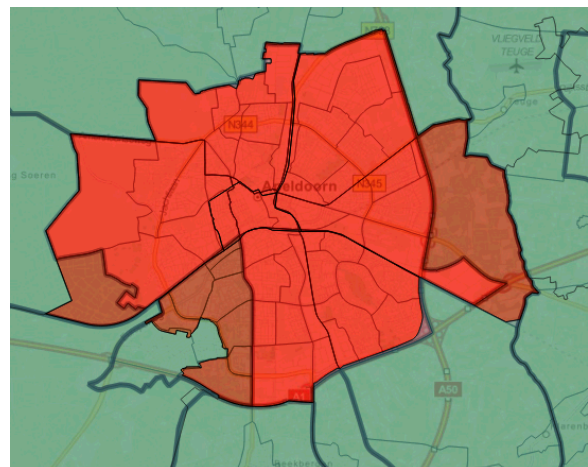
Almere



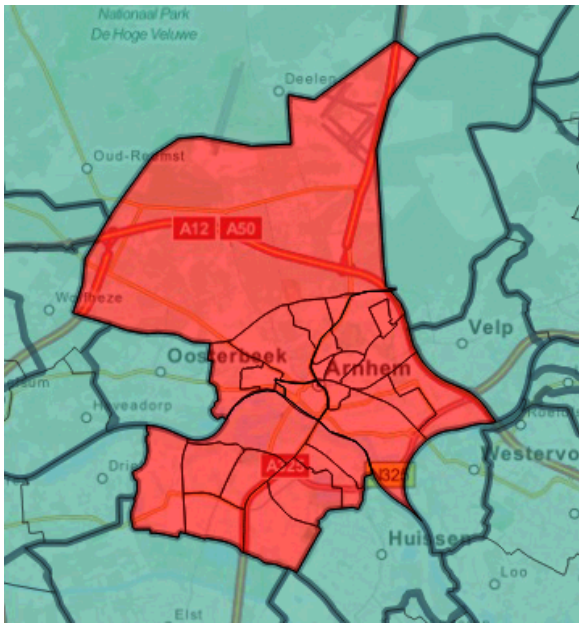
Amersfoort



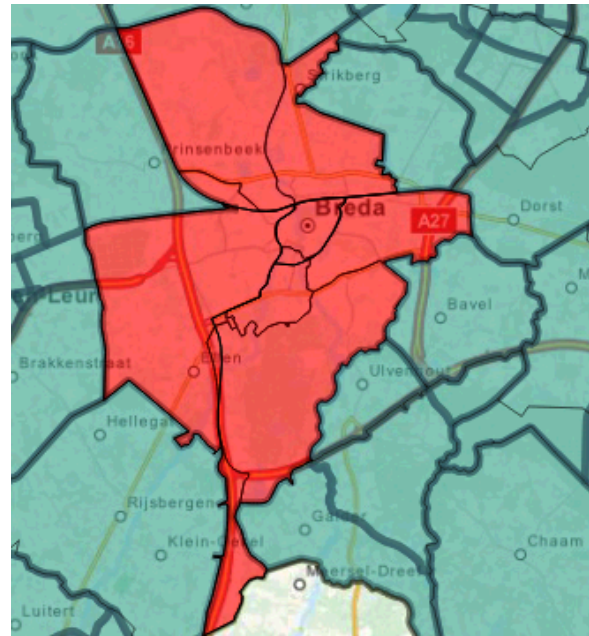
Amsterdam



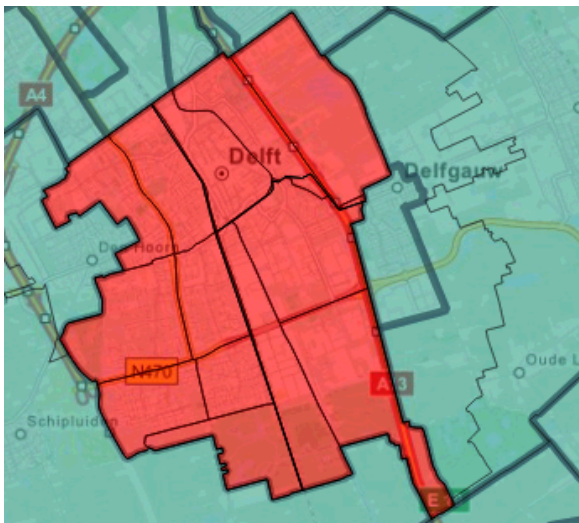
Apeldoorn



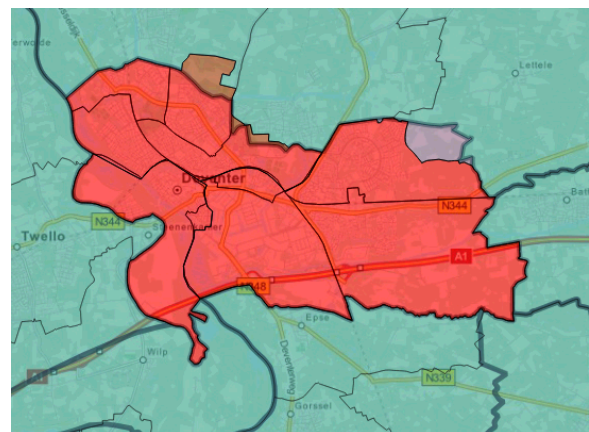
Arnhem



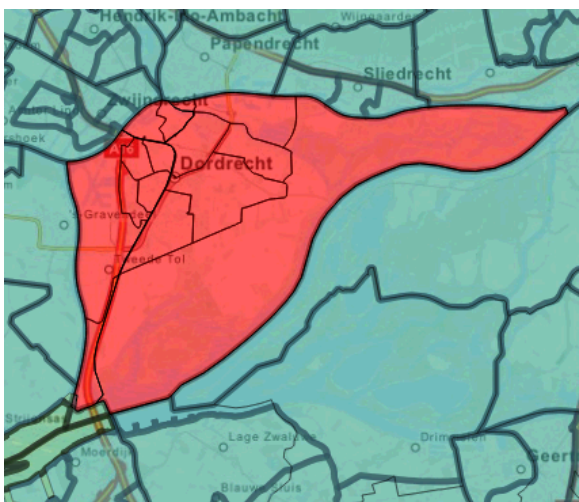
Breda



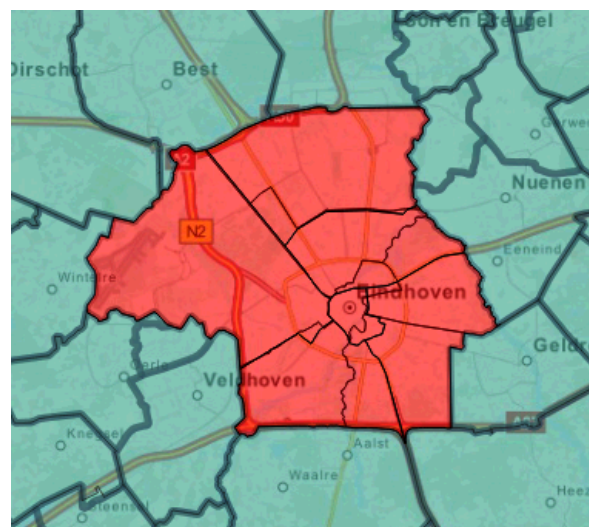
Delft



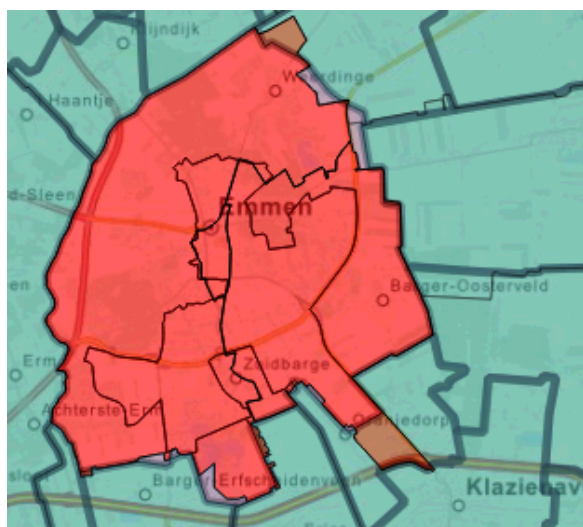
Deventer



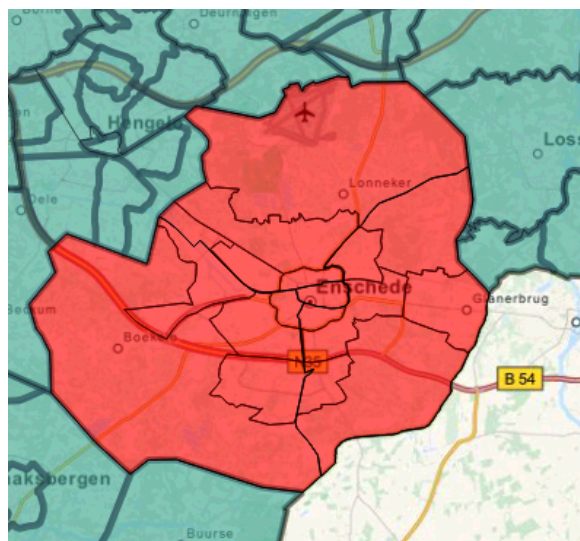
Dordrecht



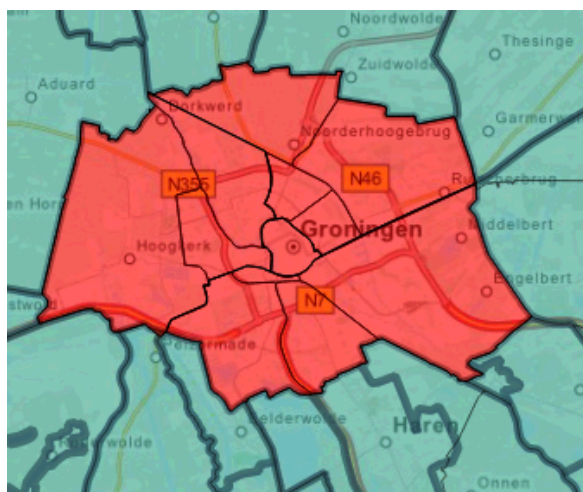
Eindhoven



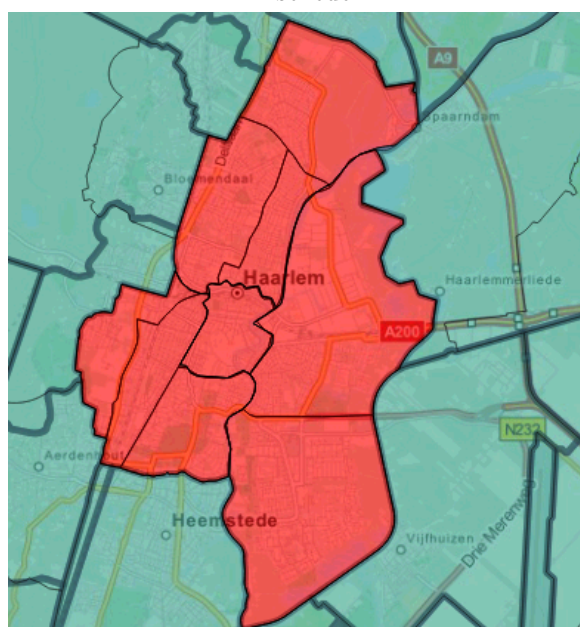
Emmen



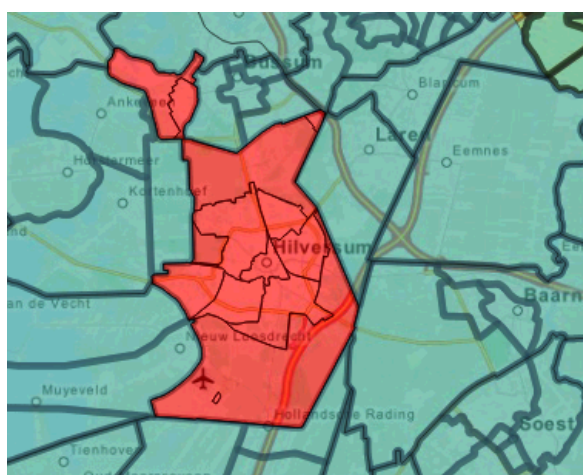
Enschede



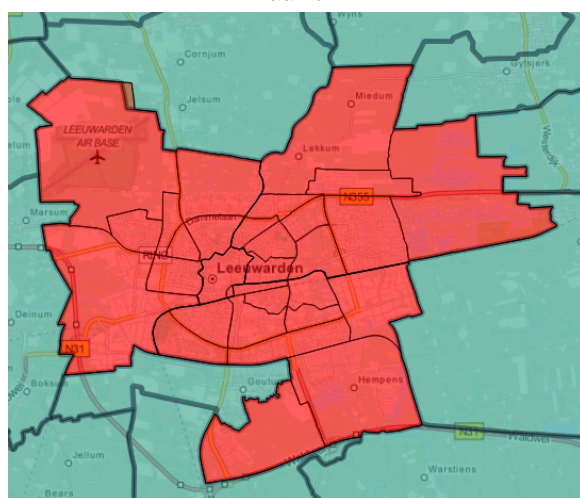
Groningen



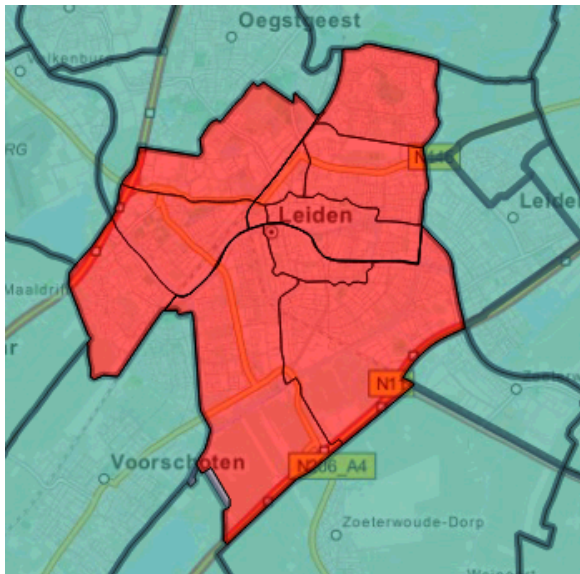
Haarlem



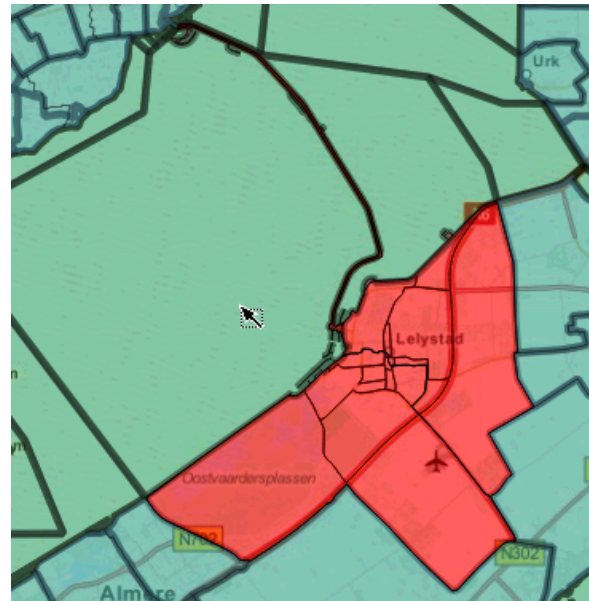
Hilversum



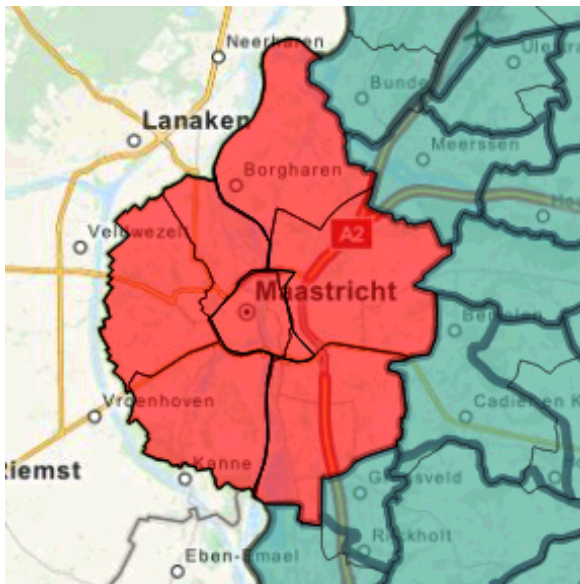
Leeuwarden



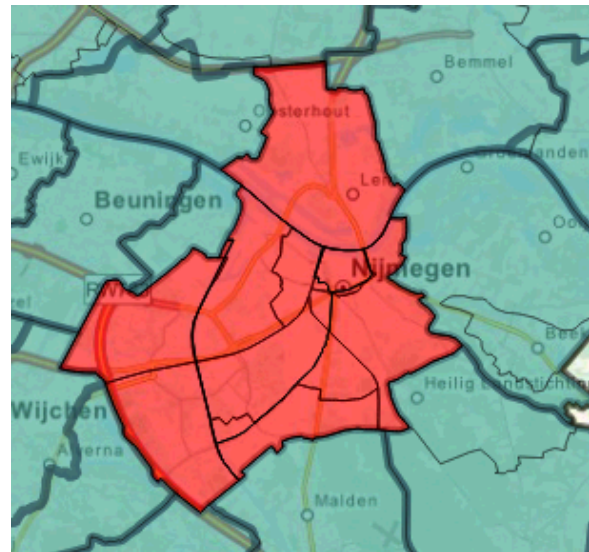
Leiden



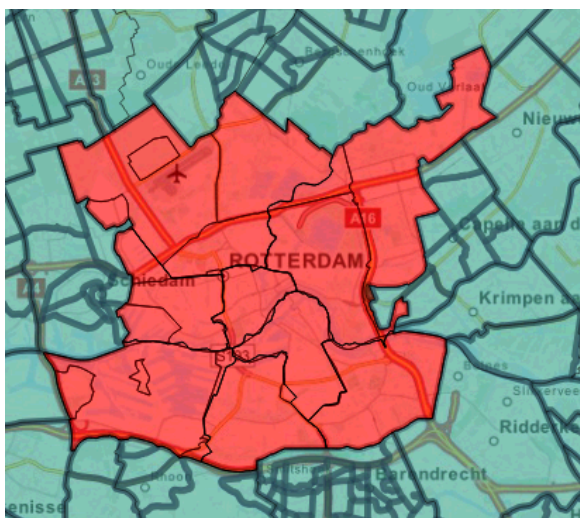
Lelystad



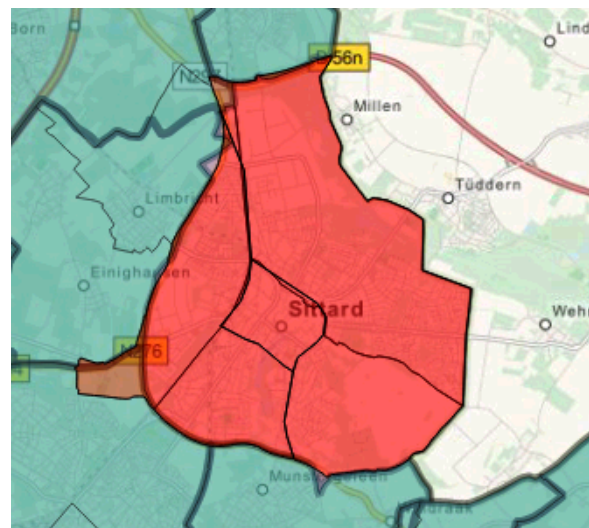
Maastricht



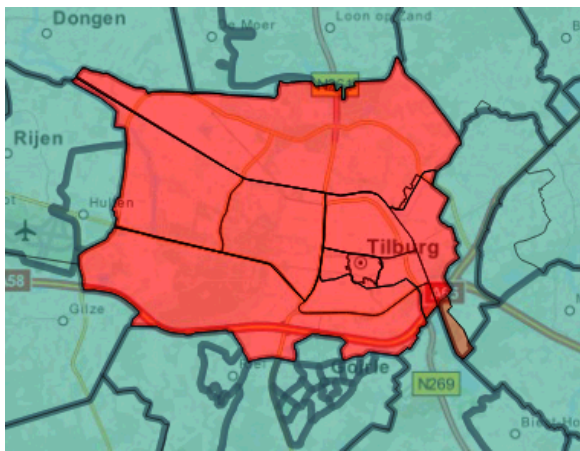
Nijmegen



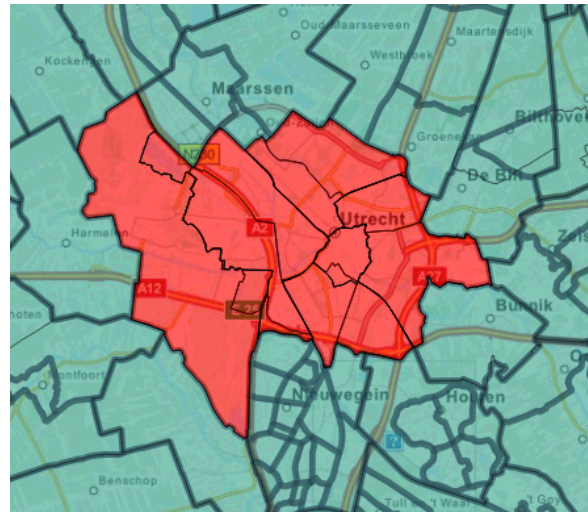
Rotterdam



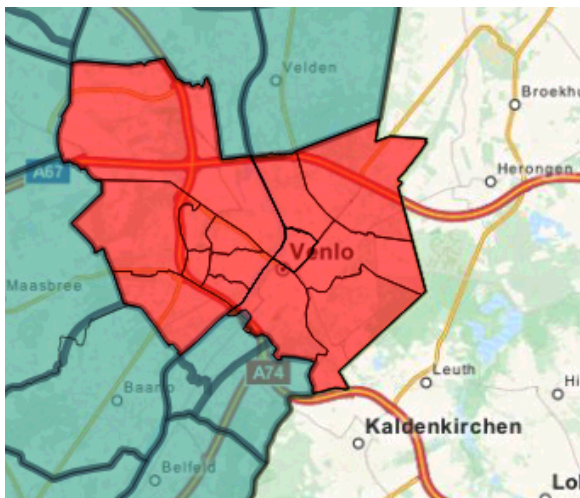
Sittard



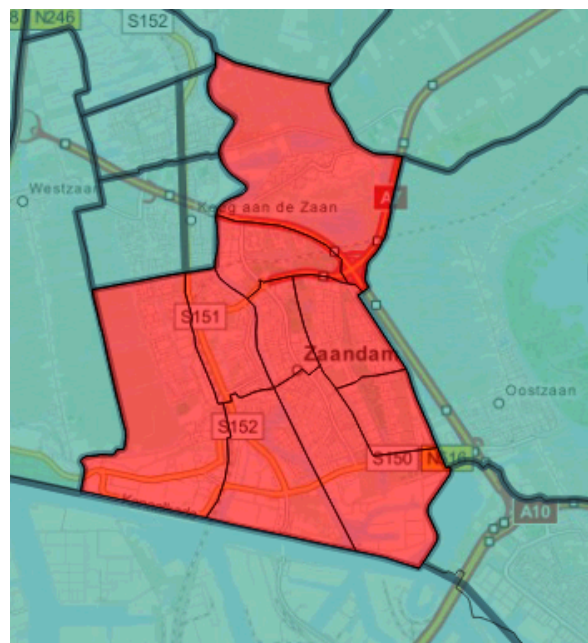
Tilburg



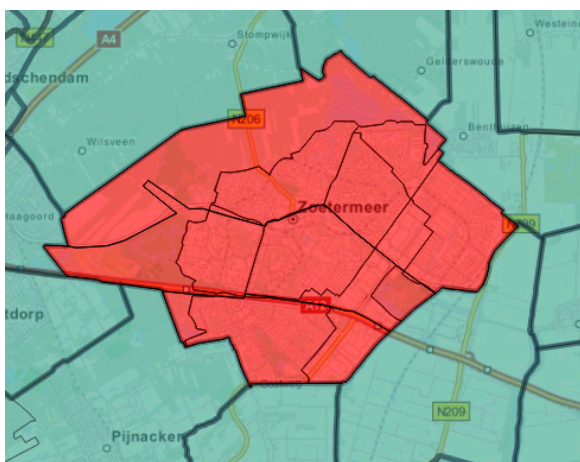
Utrecht



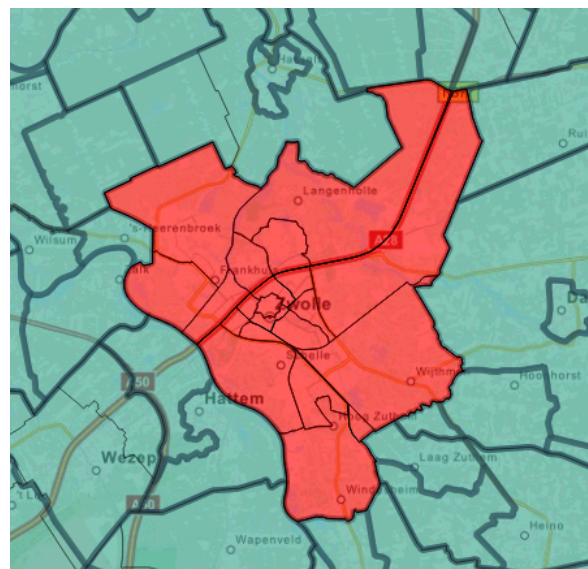
Venlo



Zaandam



Zoetermeer



Zwolle