

Identification of spatio-temporal factors affecting arrivals and departures of shared vehicles.

Master's Thesis



Technische Universität München

David Durán Rodas



TECHNICAL UNIVERSITY OF MUNICH

CHAIR OF TRANSPORTATION SYSTEMS ENGINEERING

Master's Thesis

**Identification of spatio-temporal factors
affecting arrivals and departures of shared
vehicles.**

Methodology for comparing multiple cities on a local level using
open-source data

David Durán Rodas

supervised by

Univ.-Prof.Dr. Constantinos ANTONIOU
M.Sc. Emmanouil CHANIOTAKIS

October 30, 2017

Abstract

Car and bike sharing are two categories of shared mobility that have presented environmental, economic and social benefits. Identification of their exogenous factors is needed to expand such concepts to new cities and to increase their performance and reliability. Additionally, ICT development has helped to increase the collection and sharing of data from the transport and geography sectors. However, the analysis and processing of such data has also become more difficult. Therefore, an automated methodology was formulated to correlate open-source arrivals and departure rates from shared transportation systems with exogenous factors from open geographic sources in multiple cities on a local scale. This methodology consisted of automated collection, analysis and processing of data as well as building of an automated model and selection of the most crucial variables using three methods: stepwise regression, GLM, and GBM. Daily average arrivals and departures in six cities in Germany (689 stations, 3.5Gb) using the bike sharing system "Call a Bike" were used to automatically identify the relationships with exogenous factors obtained mainly from OpenStreetMap (5.9Gb). A total of 324 models were built to correlate around 200 pre-selected independent variables with the departures and arrivals from the last 3.5 years. GBM was found to fit the validation set better, whereas stepwise regression was found to perform adequately with fewer variables than other models. An indicator of the good performance of the variables selected from the resulting models were the facts that they were logical (e.g., pubs had a high influence at night) and also that they were present in the literature.

Contents

I	Introduction and Literature Review	1
1	Introduction	2
1.1	Problem Statement	2
1.2	Need	3
1.3	Objectives and research question	3
1.4	Contributions	5
1.5	Research Framework	5
1.6	Report structure	5
2	Literature Review	8
2.1	Shared Mobility	8
2.1.1	Shared economy	8
2.1.2	Shared mobility definition and classification	8
2.1.3	Car sharing	9
2.1.4	Bike sharing	10
2.1.5	Impacts of car and bike sharing	11
2.2	Model building and selection	15
2.2.1	Model selection, assessment and validation	15
2.2.2	Multiple Linear Regression	17
2.2.3	Generalized Linear Models (GLM)	17
2.2.4	Model Diagnostics	18
2.2.5	Collinearity	19
2.2.6	Variables Selection	20
2.2.7	Decision tree methods	22
2.2.8	Gradient Boosting Machine (GBM)	23
2.3	Factors affecting the deployment of car and bike sharing	24
2.3.1	Factors affecting the deployment of car sharing systems	27
2.3.2	Factors affecting the deployment of bike sharing systems	28
II	Methodological framework	32
3	Data collection, analysis, and processing	35
3.1	Data collection	35
3.2	Time intervals	36
3.3	Zones of influence	36
3.4	Dependent variable: average daily arrivals and departures	38
3.5	Calculation of the indicators for the spatial variables	39
3.6	Pre-selection of variables	40
3.7	Exploratory data analysis (EDA)	41

4	Model building and selection	42
4.1	Detecting and addressing collinearity	42
4.2	Model building	42
4.3	Model diagnostics	46
4.4	Model assessment and selection	47
4.5	Principal variables selection	48
III	Case of Study	49
5	Area of implementation	50
5.1	Germany	50
5.2	Study Cities	50
5.3	Car sharing in Germany	52
5.4	Bike sharing in Germany	52
5.4.1	Call a bike	53
6	Results	57
6.1	Data collection, analysis, and processing	57
6.1.1	Data collection	57
6.1.2	Exploratory data analysis (I)	60
6.1.3	Time intervals	66
6.1.4	Zones of influence	67
6.1.5	Calculation of indicators	68
6.1.6	Exploratory Data Analysis (II)	69
6.2	Model building and selection	73
6.2.1	Detecting and addressing collinearity	73
6.2.2	Model building and assessment	74
6.2.3	Models assessment and selection	88
6.2.4	Principal variables selection	91
6.2.5	Performance of the regression methods using other test cities	91
6.3	Discussion of the results	98
7	Conclusions	101
7.1	Conclusions	101
7.2	Recommendations and future work	102
A	Summary of the independent variables	112
B	Analysis between the rentals and thier most correlated variables	116
C	Analysis between the rentals and the most successful factors named on the literature	118
D	Model assessment parameters	120
D.1	Stepwise regression	120
D.2	GLM + lasso	120
D.3	GBM	120
E	Fitted vs observed model values	130
E.1	Stepwise regression	130
E.2	GLM	134
E.3	GBM	138

F	Residual analysis	142
F.1	Stepwise regression	142
F.2	GLM	146
F.3	GBM	150
G	Predicted vs observed model values	154
G.1	Stepwise regression	154
G.2	GLM	158
G.3	GBM	162

List of Figures

1.1	Problem statement	4
1.2	Research framework	6
2.1	Car sharing worldwide growth (top) & number of cities employing the system (bottom) [2014]	10
2.2	Bike sharing worldwide growth (top) & number of cities employing the system (bottom) [2014]	12
2.3	Car and Bike sharing impacts	13
2.4	The number of parameters	15
2.5	Methodological framework	34
3.1	Delineation of the zones of influence in station-based shared systems	38
3.2	Calculation of indicators	40
4.1	Model building and selection	43
4.2	Example k-folds cross-validation test to find the best value of λ	45
4.3	Example k-folds cross-validation test to find the best number of iterations (trees)	46
4.4	Example of finding the λ value for the Boxcox transformation	47
5.1	Location of the cities of the study	51
5.2	Evolution of car sharing in Germany (2016)	53
5.3	Fleet size of car sharing fleet size (left) and clients (right) in Germany (2016)	54
5.4	Trip purposes for StadtRAD users (2016)	55
6.1	Call a Bike: Rentals per city (01/14 - 05/2017)	58
6.2	Location of the stations	59
6.3	Distribution and growth of the rentals in the cities of the study	60
6.4	Trips/day per 1000 inhabitants	61
6.5	Analysis of trips performed	62
6.6	Monthly rentals vs month	63
6.7	Monthly rentals vs month	63
6.8	Daily rentals vs. day of the week	64
6.9	Hourly distribution and definition of times intervals	65
6.10	Spatial demand distribution in cities of the study	66
6.11	Correlation matrices between days of the week by time interval	67
6.12	Summary of the dependent variables	71
6.13	Pearson correlation: rentals vs. variables with the highest correlation	72
6.14	Spearman's correlation: rentals vs. variables with the highest correlation	74
6.15	Pearson correlation vs spearman correlation of the rentals with the independent variables	75
6.16	Pearson correlation vs spearman correlation after a logarithmic transformation of the rentals with the independent variables	76

6.17	Example of heteroscedacity (Stepwise regression)	78
6.18	Example of homoscedacity (Stepwise regression with logarithmic transformation)	79
6.19	Example of GBM relative influence of variables ("WA1p" with BoxCox transformation)	84
6.20	Sensitivity analysis to set a threshold for the variables selection in GBM (No transformations)	85
6.21	Sensitivity analysis to set a threshold for the variables selection in GBM (Log transformation)	86
6.22	Sensitivity analysis to set a threshold for the variables selection in GBM (Boxcox transformation)	87
6.23	Comparison of the R^2 adjusted from the fitted values of different models	88
6.24	Comparison of the R^2 values from the validation of different models	89
6.25	Best validated models per regression method	89
6.26	Models comparison	90
6.27	Correlation between the different methods variables	92
6.28	Variables with most influence on the built models	93
6.29	Variables with most influence (GBM Logarithmic transformation)	95
6.30	Variables with most influence (GBM Boxcox transformation)	96
6.31	Variables with most influence (Stepwise regression Logarithmic transformation)	97
B.1	Analysis of the most correlated variables	117
C.1	Analysis between the rentals and the most successful factors named on the literature	119
E.1	Stepwise regression fitted vs observed values (No Treatment)	131
E.2	Stepwise regression fitted vs observed values (Logarithmic transformation)	132
E.3	Stepwise regression fitted vs observed values (Boxcox transformation)	133
E.4	GLM fitted vs observed values (No transformations)	135
E.5	GLM fitted vs observed values (Logarithmic transformation)	136
E.6	GLM fitted vs observed values (Boxcox transformation)	137
E.7	GBM fitted vs observed values (No transformations)	139
E.8	GBM fitted vs observed values (Logarithmic transformation)	140
E.9	GBM fitted vs observed values (Boxcox transformation)	141
F.1	Stepwise regression residuals vs fitted values (No transformations)	143
F.2	Stepwise regression residuals vs fitted values (Logarithmic transformation)	144
F.3	Stepwise regression residuals vs fitted values (Boxcox transformations)	145
F.4	GLM residuals vs fitted values (No transformations)	147
F.5	GLM residuals vs fitted values (Logarithmic transformation)	148
F.6	GLM residuals vs fitted values (Boxcox transformations)	149
F.7	gbm residuals vs fitted values (No transformations)	151
F.8	gbm residuals vs fitted values (Logarithmic transformation)	152
F.9	gbm residuals vs fitted values (Boxcox transformations)	153
G.1	Stepwise regression predicted vs observed values (No transformations)	155
G.2	Stepwise regression predicted vs observed values (Logarithmic transformation)	156
G.3	Stepwise regression predicted vs observed values (Boxcox transformation)	157
G.4	GLM predicted vs observed values (No transformations)	159
G.5	GLM predicted vs observed values (Logarithmic transformation)	160
G.6	GLM predicted vs observed values (Boxcox transformation)	161
G.7	GBM predicted vs observed values (No transformations)	163
G.8	GBM predicted vs observed values (Logarithmic transformation)	164
G.9	GBM predicted vs observed values (Boxcox transformation)	165

List of Tables

2.1	Weather variables in shared mobility studies	26
2.2	Studied exogenous factors in a literature selection	26
2.3	Time clustering types	27
2.4	Related work concerning factors affecting SBBS	29
2.5	Factors affecting the deployment of SBBS	30
2.6	Distance of influence from the stations	30
6.1	Nomenclature for the different dependent variables	68
6.2	Sensitivity analysis to chose the buffer distance	69
6.3	Sensitivity analysis to choose the standard deviation threshold	69
6.4	Summary of the dependent variables	70
6.5	Pearson's correlation Attractions vs Productions	72
6.6	Comparison of the performance of transformation of the dependent variable	73
6.7	Comparison of the techniques to address collinearity	74
6.8	Sensitivity analysis to choose the best way to address collinearity	77
6.9	Comparison of the results from the Stepwise regression	78
6.10	Stepwise regression results (No transformation)	80
6.11	Stepwise regression (Logarithmic transformation)	81
6.12	Stepwise regression results (Boxcox transformation)	82
6.13	Comparison of the results from the GLM regression	83
6.14	Comparison of the results from the GBM regression	83
6.15	R^2 from validation by testing other cities (GBM with log transformation)	94
A.1	Summary of the independent variables	112
D.1	Results from stepwise regression (No transformations)	121
D.2	Results from stepwise regression (Log transformation)	122
D.3	Results from stepwise regression (BoxCox transformation)	123
D.4	Results from GLM (No transformations)	124
D.5	Results from GLM (Logarithmic transformation)	125
D.6	Results from GLM (Boxcox transformation)	126
D.7	Results from GBM (No transformations)	127
D.8	Results from GBM (Logarithmic transformation)	128
D.9	Results from GBM (Boxcox transformation)	129

List of Abbreviations

AIC	Akaike information criterion
BIC	Bayesian information criterion
BS	Bike sharing
CBD	Central business district
CS	Car sharing
DBSCAN	Density-based spatial clustering of applications with noise
DE	Direct elimination
EDA	Exploratory data analysis
FF	Free-floating-based
FFBS	Free-floating-based bike sharing
FFCS	Free-floating-based car sharing
GLM	Generalized linear models
GB	Gigabits
GBM	Gradient boosting machine
ICT	information and telecommunication technologies
lasso	Least absolute shrinkage and selection operator
LL	Log-Likelihood
log	Logarithm
MSE	Mean squares error
OSM	OpenStreetMap
OLS	Ordinary least squares
P2P	Peer to peer
POI	Point of Interest
R	R statistical programming language
SD	Standard Deviation
SB	Station-based
SBBS	Station-based bike sharing
SBCS	Station-based car sharing
SSE	Sum of squared errors
SSR	Sum of squared residuals
VIF	Variable Inflation Factor
VKT	Vehicle kilometers traveled
VMT	VMT Vehicle miles traveled

Acknowledgements

I would like to thank my supervisors Prof. Dr. Constantinos Antoniou and MSc. Emmanouil Chaniotakis, who guided me in an extraordinary way through this thesis.

Secondly, I would like to thank and dedicate this thesis to my wife, family, and friends who supported me during my studies.

Finally, I would like to thank the government of Ecuador for funding my Master's studies through the "Secretaria de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT)"

Part I

Introduction and Literature Review

Chapter 1

Introduction

This first chapter introduces the thesis. It starts pointing the problem statement including the adversary effects on transportation and the other hand, the development of ICT on the data and the transport sector. Then, the need to develop shared mobility and identify their influencing factors is explained leads to the objectives and research question of the thesis. Finally, the contributions of the thesis, its framework and the structure of the research are summarized.

1.1 Problem Statement

Private transportation trips have shown in Europe a significant growth, and they are expected to continue growing (EC, 2016). This growth has presented environmental, economic, and social impacts. Transportation is considered a major source of emissions at global and local levels, particularly in urban areas contributing in problems as climate change, air, water and soil pollution and noise. Road transport generates around 23% of the total CO_2 emitted in Europe, and, the 39% of the total amount of NOx emissions and 13% of the total particulate matter (EC, 2016). Besides, environmental issues, transportation has provoked additional economic and social problems, such as congestion, accidents, health problems, mobility gaps, land consumption among others (Rodrigue et al., 2016). Congestion costs each year in Europe about one percent of the gross domestic product (EC, 2016). The demand for parking places creates space consumption problems mainly in central areas since cars spend most of the time parked (Rodrigue et al., 2016). However, private transport sometimes is required to supply some specific needs in areas where public transport does not provide the required travel routes or restricts the transportation of goods. Therefore, in these cases, there is limited transport access for low-income that cannot use a private vehicle. This social exclusion might impede access to services, leisure activities, education, or employment (Mackett and Thoreau, 2015).

On the other hand, the development of information and telecommunication technologies (ICT) has allowed mainly public and collaborative organizations to collect and publish large databases as open-source data, i.e., non-privacy-restricted and nonconfidential data making it available without restrictions (Janssen et al., 2012). The availability of open-source data has grown significantly mainly in domains from traffic, weather, geographical, tourist, among others. For instance, open-source data have presented relative high precision and actualized geographical information. Another example is in the transport sector, real-time data collection of vehicles' location is available from GPS systems installed in the vehicles. However, these large databases have presented issues related to their analyzing and process.

Additionally, ICT has presented a potential to reduce negative transport's impacts by allowing a new type of mobility services. The main enabler of these technologies has been the broad connectivity that ICT offers and that includes handheld devices that are connected and allow for better coordination concerning the use of shared systems.

1.2 Need

There is a need to develop sustainable transportation systems because of global climate change, oil dependency, and congestion derived from the transportation sector (Shaheen et al., 2012). Sustainable mobility will ensure the "mobility of people and goods concerning energy, environment, safety and security as well as socio-economic issues" (Nowicka, 2016). This approach requires actions to reduce the need to travel by private cars and the trips' length, and increase modal shift to active modes and efficiency of the transportation systems (Banister, 2008).

One sustainable mobility concept is shared mobility (Shaheen et al., 2010a; Shaheen and Cohen, 2012). Shared mobility is part of the movement of the shared economy; which essentially prioritizes utilization over ownership of goods (Cheng, 2016). In this concept, shared mobility prioritizes the utilization of vehicle use over its ownership. Although shared mobility does not eliminate the external effects of transport, it is a partial solution to mitigate the problems of emissions, congestion, restricted available parking space in urban areas, and access exclusion (see Literature Review). The two prevailing forms of shared mobility are car sharing (CS) and bike sharing (BS). They are defined as the shared use of a car or a bicycle's fleet (Shaheen et al., 2010a). In such shared transportation systems, a user accesses a fleet of shared-vehicles (car or bike) by joining an entity that maintains the fleet of vehicles by usually paying a fee for the usage of the shared transport mode (Shaheen et al., 2010b).

In the last years, shared mobility projects have grown in around 500 cities around the world (Cohen and Muñoz, 2016). Therefore, there is a need to define a robust methodology that would allow the identification of the factors affecting the use of these shared modes of transport in a way that would allow for prediction of their demand in a different setting and with only open-source data. To identify of their success factors are important for mainly five reasons (Kortum et al., 2016):

- To assist operators and policymakers on their deployment of shared transport modes.
- To increase the reliability of implementations and policies
- To reduce the risk of supply-demand imbalance in existing systems.
- To expand these transport modes to new business areas and other cities.
- To model shared mobility as a special case of future shared autonomous vehicles (Schmöller et al., 2015).

So far and to the best of the author's knowledge, the state-of-the-art does not provide a consistent methodology that takes into account multiple cities to identify the correlation between exogenous factors affecting the demand for shared vehicles at a local level using open-source data.

Therefore, this thesis explored an automated methodology to collect and analyze open-source arrivals and departures from shared vehicles and also open-source exogenous factors from multiple cities and correlate them in a local level over time (see Figure 1.1). This methodology included the selection of a regression model that fitted better the data and also a ranking list of the variables the influences more the demand of shared vehicles.

1.3 Objectives and research question

This study is motivated to contribute to the continued development of shared mobility. Shared transportation systems can contribute reducing environmental, congestion, and access and space problems of private vehicle ownership. Thus, to enhance their demand and the reliability of their implementation, the interrelation between shared vehicles and the exogenous factors which affect them over time must to be analyzed. Therefore, the main objectives of this research are:

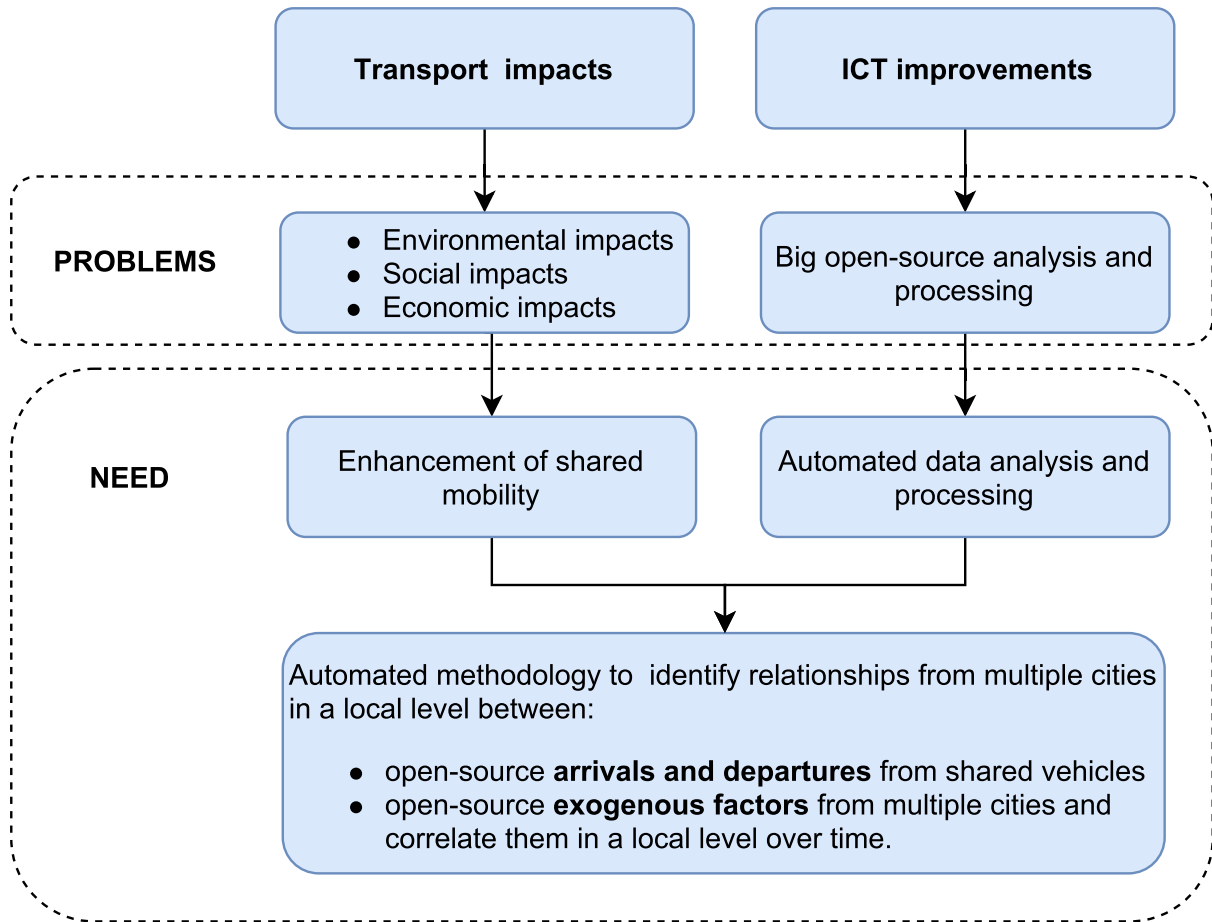


Figure 1.1: Problem statement

1. To develop an automated methodology to identify relationships between exogenous factors affecting the demand for shared vehicles taking into account multiple cities on the local scale over time using open-source data.
2. To build models using exogenous factors that fit the historical arrivals and departures of shared vehicles as best as possible across in multiple cities over time.
3. To rank the most influential exogenous variables affecting the deployment of shared vehicles.
4. To evaluate the possibility to perform a shared mobility research study based only on open-source transport-related data.

To achieve these objectives, the following research question has to be answered:

- *How do exogenous factors influence the demand of shared vehicles at a local scale over time?*

Hypothesis: The factors that are expected to influence the demand for shared vehicles are:

1. Points of interest (universities and leisure activities (beer garden, cinema, pubs)),
2. Land-use (proximity to residential areas, green areas, water areas)
3. Transport infrastructure (density of transit stations, residential streets, and cycling infrastructure).

1.4 Contributions

This thesis contributes on a theoretical, methodological and practical level:

- Theoretical contributions
 1. Synthesis of the impacts of car and bike sharing.
 2. Summary of main factors affecting shared vehicles in multiple cities on a city level or in a city on a zone of influence level.
 3. Potential future research.
- Methodological contributions The methodological contributions apply a method to:
 1. Automated processing and selecting open-source data and implementing them in a model.
 2. Automated calculation of relevant indicators for different spatial variables.
 3. Automated process for building models for different time intervals that correlates arrivals and departures of shared vehicles in a zone of influence scale for multiple cities.
 4. Automated process for ranking most relevant exogenous variables that influence the arrivals and departures of shared vehicles in multiple cities on a zone of influence level over time.
- Practical contributions
 1. A ranking of the most relevant exogenous variables that influence the arrivals and departures of bike sharing systems in six cities in Germany on a zone of influence level over time.
 2. Models for different time units that correlated the existing infrastructure with the demand of the bike sharing systems in six cities in Germany at a zone of influence level.

1.5 Research Framework

A framework for this research was developed to proceed through systematically (see Figure 1.2). It contains seven main sections, starting with an introduction (Chapter 1), and followed by the literature review (Chapter 2), the methodological framework (Chapter 3,4), the area of implementation (Chapter 5), the (Chapter 6), and discussion and conclusions (Chapter 7).

The literature review includes two main topics: shared mobility concepts and modeling theory (Chapter 2). After these concepts are explained, the related work of factors influencing demand of shared vehicles is described. With those background concepts, a methodology was established beginning with the data collection, analysis, and process (Chapter 3). The construction of the model is explained and how they and the most influencing variables were selected (Chapter 3). The area of study is described (Chapter 5) and then the methodological framework is implemented in this area (Chapter 6). The results of the implementation are shown (Chapter 6) and discussed to finally, present conclusions and recommended further research (Chapter 7).

1.6 Report structure

This report is split into three parts, followed by the discussion and conclusions.

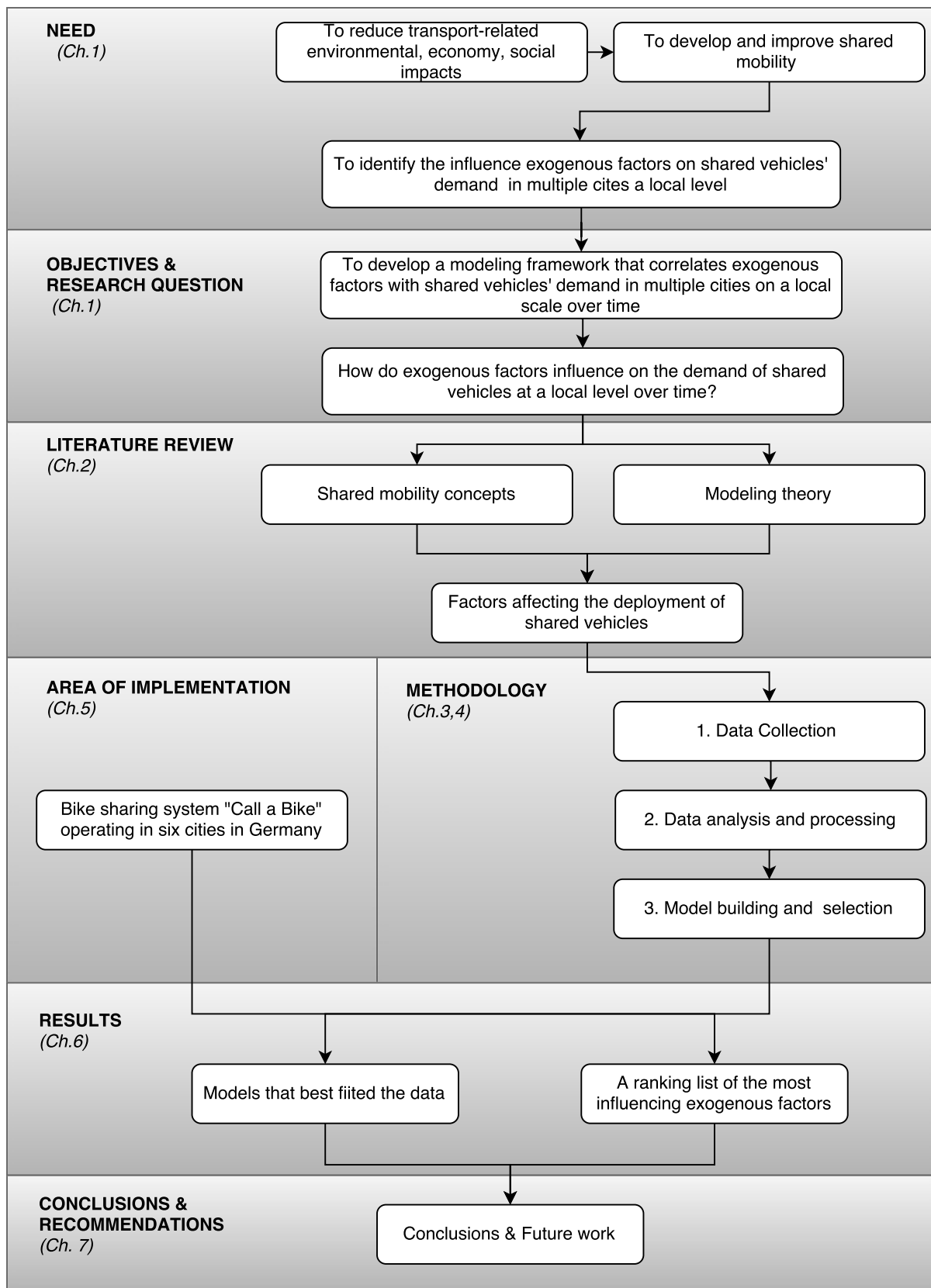


Figure 1.2: Research framework

Part I: Introduction and Literature This part introduces the study (Chapter 1), followed by a literature reviewed in Chapter 2. As the need to identify exogenous factors affecting

the demand for shared vehicles, shared mobility's state-of-art is described in Chapter 2 including concepts, trends, and main impacts. Main modeling theory is also explained considering different types of models construction, variables selection techniques and diagnostics and assessment methods. This chapter concludes with the related work on identifying factors influencing the deployment of shared vehicles.

Part II: Methodological Framework The second part indicates the methodology to achieve the objectives of the research. Chapter 3 begins by explaining how and which data should be collected. It follows describing how the data should be analyzed and then process to get the input variables for the building of the regression models. Then, Chapter 4 shows the methodology to construct three types of models and how to assess them to obtain the model that best fit the dataset and the variables that influence the most.

Part III: Case of Study The third part guides the reader through the implementation of the methodology in six cities in Germany using the bike sharing system "Call a Bike". First, Chapter 5 gives an overview of Germany and these cities including an outline of the mentioned bike sharing system. Then, Chapter 6 shows the implementation of the methodology and their results.

Discussion and conclusion . The last chapter is dedicated to summarizing and discussed the results, and also present the conclusions of the research and future recommended work.

Chapter 2

Literature Review

The literature review was split into three parts. First, the main shared mobility concepts were described focusing on car and bike sharing. It also included their demand trends and their main environmental, economic and social impacts. Then, the theory of modeling was described focused on linear and nonlinear model building and selection. Finally, the related work on models that identified the influencing factors on car and bike sharing were summarized

2.1 Shared Mobility

This section introduces the main concepts of shared economy to understand shared mobility. Then, shared mobility is defined and its main types are summarized. It emphasized car and bike sharing concepts, their history, demand trends and principal impacts in an attempt to specify the exact properties that make the system viable and to start identifying their influencing factors.

2.1.1 Shared economy

Sharing items through online services have increased in the past years (Böckmann, 2013). Based on this fact, a new term "shared economy" was born in the scientific literature. Shared economy is a social-economic phenomenon that prioritizes utilization over ownership. Essentially, individuals or organizations aim at utilizing regarding time used their belongings. Prominent examples of such under-utilization are usually met in many sectors, to name but a few, they can range from cars, accommodations and even household articles (Böckmann, 2013). As an example for accommodations, a private apartment is idle 23 hours per day on average, while houses are also empty when owners go for vacation or are just out of town. This under-utilization has been the driving force for the development of many successful business models that are essentially categorized as shared economy business models (Cheng, 2016).

The main reasons behind their successful implementation are social, technical and economic components (Böckmann, 2013). Social drivers are mainly the high population concentration, which facilitates sharing; environmental awareness since less raw materials are used, and the wiliness to have a collaborative society. The technological component includes recent advances in technology such as the ease of online access services, the efficiency of social networks, mobile devices, and payment systems. Finally, the economic factors deal with individual savings under certain circumstances of sporadic use and also an ease to access luxury articles (e.g., luxury means of transport) (Cheng, 2016; Böckmann, 2013)

2.1.2 Shared mobility definition and classification

Shared mobility is a component of the shared economy. It is defined as the "shared use of a vehicle." It allows the user to get access to a private transport mode "on an as-needed" basis"

(Shaheen et al., 2015). It can be classified into seven categories:

1. Car sharing: Roundtrip, one way, peer to peer (P2P).
2. Bike Sharing: Public bike sharing, P2P, campus bike sharing.
3. Scooter sharing
4. Ride sharing: Car pooling
5. Alternative transit services; Shuttles
6. Courier Network services: P2P delivery services
7. On-Demand ride services: Ridesourcing, ridesplitting

The three first categories, when not a peer to peer service, are offered in the public space for a (car, bike, scooter) short-term rental (Büttner and Petersen, 2011). Users usually have to join an organization that maintains this fleet and usually pay a fee for the vehicle's usage (Shaheen et al., 2012; Shaheen and Cohen, 2012). A categorization of these systems can be defined by the use of stations and can be distinguished in three commonly seen types: (a) station-based (SB), b) free-floating (FF) and c) a mix of the two (Firnkor and Shaheen, 2016). Station based systems are centered around the fact that the start and end of a trip occurs at a station. These systems can be further categorized into round trip SB systems (return at the same station) and one-way trip SB systems (return vehicles at a different station). Free-floating is a GPS based system without fixed stations where the users can start or end their trip anywhere in a permissible area within the city. A common variant is the fusion of free-floating with station-based one-way systems (Firnkor and Shaheen, 2016).

Ride sharing considers sharing a vehicle with passengers that have the same origin and destination. Shuttles are vehicles that allow sharing rides connecting passengers to the public transport. P2P Delivery Services is a system that enables members to use a private vehicle to conduct a delivery. Ridesourcing uses apps to connect drivers with passengers. Ridesplitting involves ridesourcing that with users that take a similar route (e.g. Lyft Line and UberPOOL) (Shaheen et al., 2015).

2.1.3 Car sharing

Car sharing is a short-term rental of a car (Shaheen, 2016). It is based on the usage of a private car without the costs and responsibilities of owning one (Shaheen et al., 2015). The first experience with station based car sharing (SBCS) started in Zurich, Switzerland in 1948. Individuals, who could not buy a vehicle, shared one. In North America, the car sharing programs started in 1983 in San Francisco California. Car sharing was popularized in Berlin-Germany in 1987 and Zurich in 1988. Nowadays, car sharing technologies have spread all over the globe in five continents (see Figure 2.1)(Shaheen and Cohen, 2012). New technologies facilitate the car sharing system's usage, allowing the development of new forms, such as free-floating car sharing (FFCS).

FFCS allows users to rent and return cars at any location within the city limits. In contrast with station based car-sharing (SBCS), FFCS does not have fixed stations, and usually, the trips are one-way without a booking requirement. Available vehicles' location, auto cleanliness, and fuel level are real-time information, which is reachable by mobile phones application, calling a hot-line or online(Firnkor and Müller, 2011). The operator usually defines an area where the cars can travel, often at the city center. The renting cost usually includes a starting fee and then a time-dependent cost (Weigl and Bogenberger, 2013). The first free-floating system in operation was in Ulm, Germany in April 2009 with the operator Daimler (car2go), while in North America the first system was in Austin, the USA in May 2010 (Firnkor, 2012). The

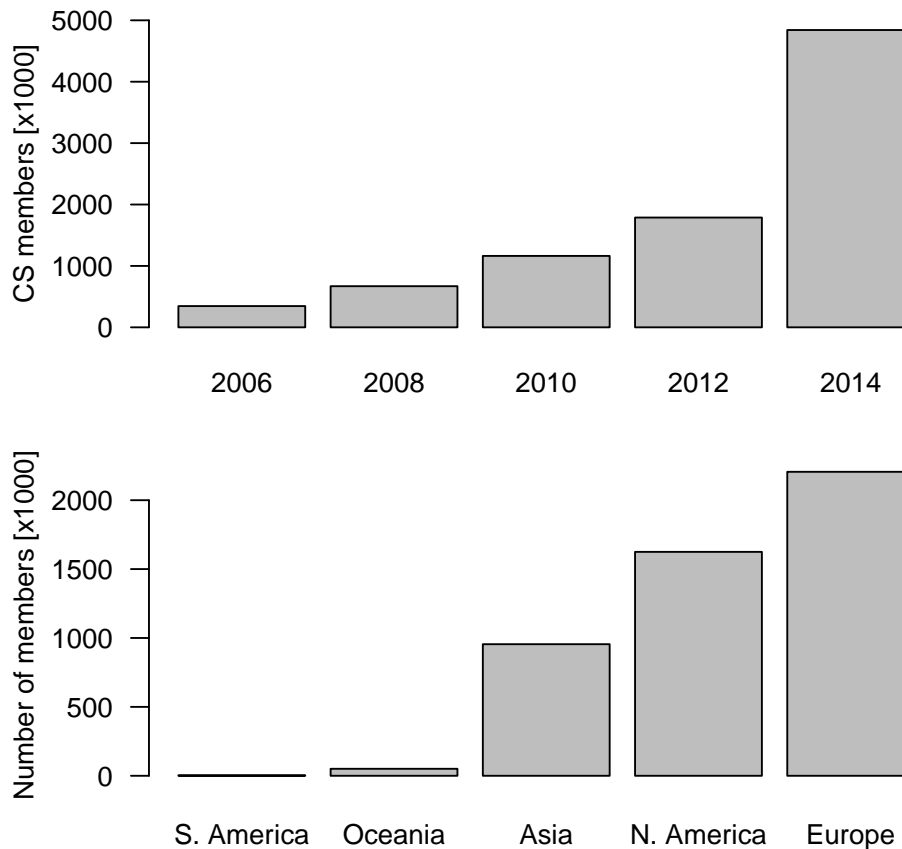


Figure 2.1: Car sharing worldwide growth (top) & number of cities employing the system (bottom) [2014]

Source: [Shaheen and Cohen \(2016\)](#)

number of memberships and the use of FFCS is growing in most of the cities, which implemented these transportation systems ([Kortum et al., 2016](#)).

Car-sharing users tend not to drive cars frequently, and they use this services for leisure purposes, shopping and to transport goods or people ([Kopp et al., 2015](#)). They can choose where is the destination of their, but not the origin because it depends on the location and availability of the vehicles. That is the reason why is challenging to estimate the origin of the trip. However, if an area has several destination trips, this means that there were several starting points ([Willing et al., 2017](#)). The disadvantage of the system is the unbalanced availability of the vehicles. Some areas have lower demand (cold spots), where the cars get stuck, but those cars are needed in areas with higher demand(hot spots) ([Weikl and Bogenberger, 2013](#)). In contrast with bikes, reallocating cars is very expensive because they have to be relocated individually or in an expensive car transporter ([Weikl and Bogenberger, 2013](#)).

2.1.4 Bike sharing

Bike sharing presents several names in literature as: "Bicycle sharing", "Bike share", "Public bicycle", "Public bike", "Public bike sharing".([Fishman et al., 2013](#)). Public bike sharing exists since 1965, and it has been present in four different generation. The first generation started in Amsterdam with fifty free and unlocked bicycles for public use. Theft and vandalism led to a coin-deposit system (second generation). Docks were implements and the deposit was not more than \$4. The first system of this type was implemented in 1995 in Copenhagen, Denmark.

However, this system collapsed mainly because of the user's anonymity. IT development helped to shape the third generation. This system includes wireless pick-up, drop-off, and bicycle tracking. In Rennes, France this system was first implemented in 1998 with a smart card to rent the bicycles. Finally, the fourth generation does not include docks. Bikes have included a GPS for real-time tracking, an on-board computer, and internet, there is integration with public transport, and the rental is automatic (Shaheen et al., 2012).

Nowadays, the high amount of research indicates the boom of bike sharing (Zhao et al., 2014). It has emerged in the major cities all over the world. There are more than 800 programs around the world with a fleet of more than 900.000 bicycles. The biggest is in Hangzhou, Paris, London and Washington D.C. (Gauthier et al., 2013).

In general, the trip propose for bike sharing is to commute to work on weekdays and for leisure and social purposes (Fishman et al., 2013)

Because of the new technologies, in free-floating bike sharing (FFBS) bikes can be locked to an ordinary bike frame, without the need for fixed stations. In contrast with Station Based bike sharing (SBBS), FFBS avoids the cost of docking stations. Thanks to their installed GPS. This transportation system can be tracked in real time allowing a smart management and reduced probabilities of bicycles theft. FFBS is more convenient for users than SBBS because the average walking distance to their destination is shorter and they do not have to worry about the bike's storage in a docking station (Pal and Zhang, 2017). However, on daily operations, the location of bicycles might be skewed leading to an unsatisfactory service. Therefore, operators should balance the distribution of bicycles. The rebalancing process is usually carried at nights when the bookings' demand is lower. In some cases, users' intervention is considered for the rebalancing. The objective of this problem is "to minimize the financial and environmental costs of re-balancing." This issue is more difficult in FFBS than in SBBS, since the nodes of SBBS are the stations and operators have to move from stations with a surplus of bikes to stations with deficit of them (Pal and Zhang, 2017).

From 2007 bike sharing systems have been growing strongly worldwide (see Figure 2.2). Europe and Asia are the continents with the most bike sharing systems worldwide (Meddin and DeMaio, 2015). In 2015, China had the biggest fleet in the world with 753.508 bicycles followed by France with 42.930 and Spain with 25.084. The biggest fleets are in Wuhan Hangzhou and Taiyuan in Asia, London, Paris, and Barcelona in Europe, and in New York in America (Meddin and DeMaio, 2015).

2.1.5 Impacts of car and bike sharing

Car and bike sharing benefits can be classified as environmental, social, and economic because of their transport and land-use-related impacts (Shaheen and Cohen, 2012; Shaheen et al., 2012; Firnkorn and Müller, 2011). The main benefit is the reduction of private cars ownership. Because of this decrease, the positive consequences are fewer vehicle kilometers traveled, individual savings, efficient use of roads and infrastructure, less raw materials consumption, and some trips change to alternative transport modes. The final benefits are fewer emissions, congestion, less space consumption, an enhancement of the image of the city and even health benefits. Figure 2.3 shows the major impacts of car and bike sharing and how they are connected to each other. The source of this figure is from studies that explained these impacts:

Reduction in private cars ownership Some users have sold their vehicles after joining a car sharing program (15.6 to 34%), or users have avoided a car purchase (e.g., U.S.: 29-68%) (Shaheen and Cohen, 2012). Additionally, Martin et al. (2010) stated that users after joining a car sharing program have an average reduction of vehicles per household from 0.24 to 0.47. Also, (Shaheen and Chan, 2015) concluded that 5.5% of bike sharing users sold or postponed a vehicle purchase. Reduction in vehicle ownership led to a decrease in the vehicles kilometer traveled (VKT), reduce congestion, reduce parking demand and

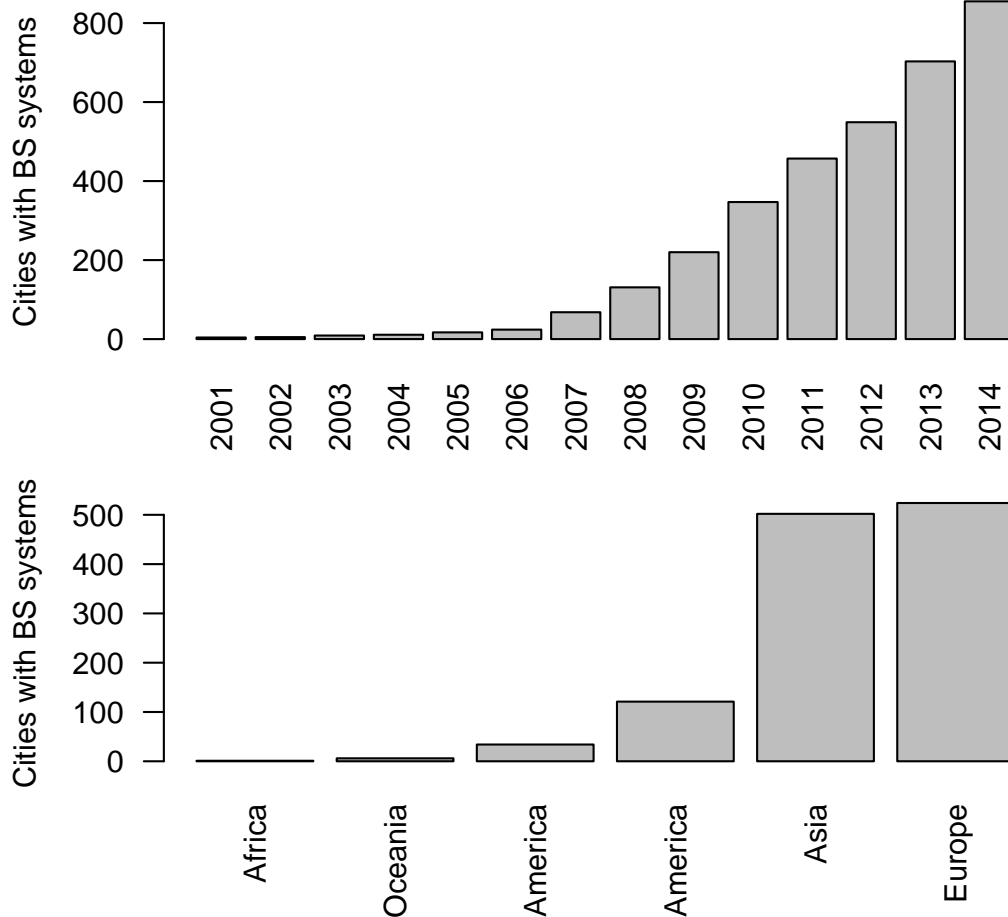


Figure 2.2: Bike sharing worldwide growth (top) & number of cities employing the system (bottom) [2014]

Source: [Meddin and DeMaio \(2015\)](#)

increase the use of public transport and active modes ([Shaheen and Cohen, 2012](#)). [Giesel and Nobis \(2016\)](#) examined the reduction of car ownership led by FFCS (DriveNow and Flinkster) after an online survey in Germany. 72% of the Flinkster users and 43% of the DriveNow users do not own a car in the household. These values are higher than the average in both cities (Berlin: 41%, Munich: 39%). Their main reasons for not owning a car are that car sharing is sufficient, and due to the ownership's costs. The principal argument for the users to give away a car would be that car sharing would always be available. Finally, those who sold their car because of joining a car sharing program correspond to 6,5% of DriveNow users and 15,3% of Flinkster users, while 1,8% and 1,7% respectively are planning to sell their cars due to car sharing.

Reduction of Vehicle Miles Traveled (VMT) / Vehicle Kilometer Traveled (VKT). ([Shaheen and Cohen, 2012](#)) calculated an average reduction of 44 % in VMT per car sharing user across North American studies. The operator Communauto estimated a decrease of 2900 km traveled per year on car sharing users ([Shaheen et al., 2010b](#)). In general, [Shaheen and Chan \(2015\)](#) concluded that car sharing reduces from 23% to 43% less VMT per year. Significant reductions in VKT are shown after the small proportion of car sharing users who gained car access, concerning those who decided to share a vehicle ([Kent, 2014](#)). Also,

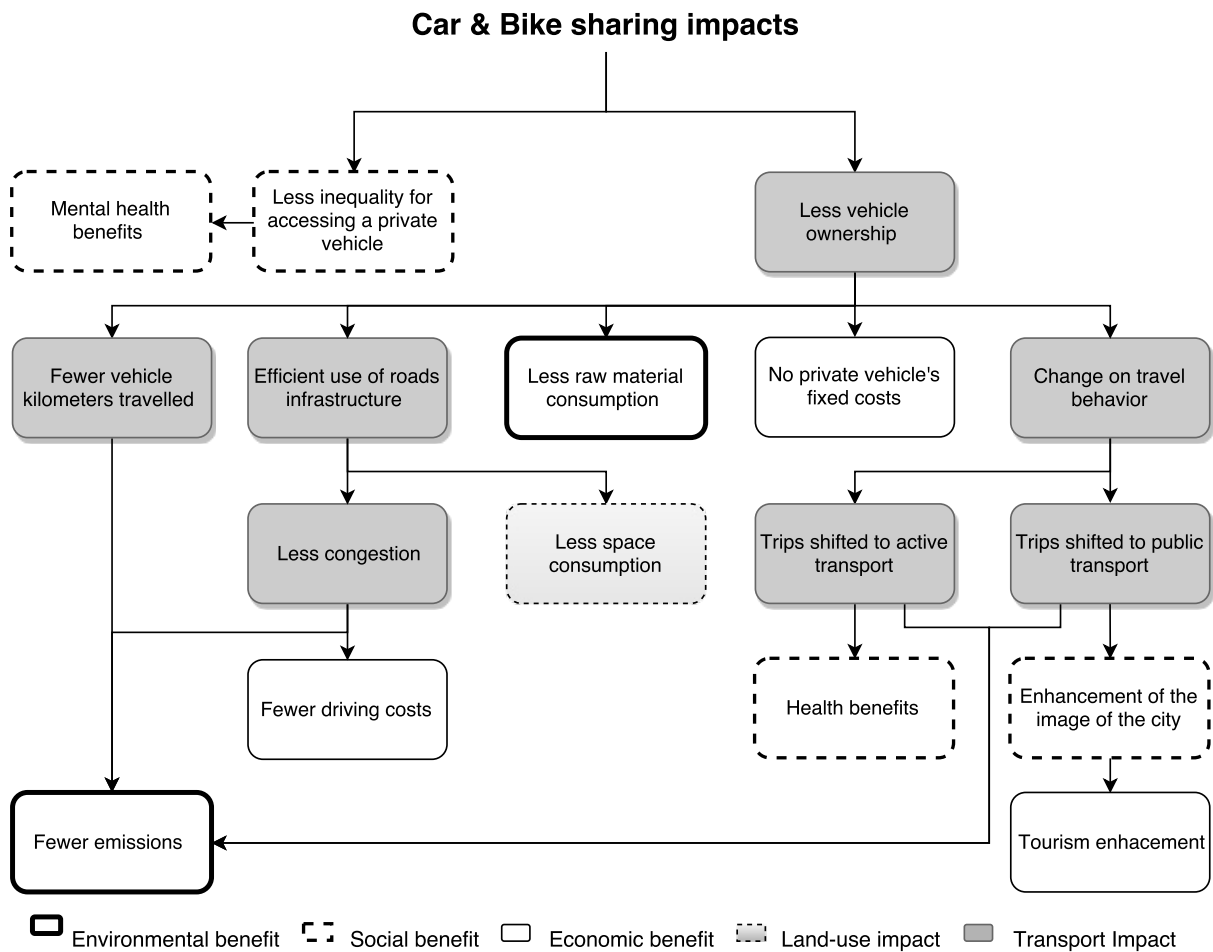


Figure 2.3: Car and Bike sharing impacts

bike sharing has replaced 7 to 20% the use of private vehicles (Shaheen et al., 2010a). In 2012, the usage of bike sharing provoked a reduction of 115.000 kilometers in Melbourne, 243.000 in Washington D.C. and 630000 km in London (Fishman et al., 2014). Around 50% of bike sharing users reduced their private car usage (Shaheen and Chan, 2015). For example, each trip in bike sharing avoids 2 to 4 kilometers driven by car (Gauthier et al., 2013). Additionally, 40% respondents of bike sharing users of a survey in North America felt they drove less private cars and 46% fewer taxi driving (Shaheen et al., 2012).

Lower GHG emissions A reduction of VMT lowers greenhouse gases (GHG) emissions (Shaheen and Cohen, 2012). Martin and Shaheen (2011) stated that car sharing reduces GHG in a range between 15% and 40% (109000 to 155 000 t GHG/year). In Europe is expected a reduction from 40 to 50% of the average of CO₂ emissions produced by car sharing users (Shaheen et al., 2010b). Also, (Shaheen and Chan, 2015) pointed that car sharing has shown a reduction of GHG emission for one household from 0.58- 0.83 metric tons per year. This quantity means a reduction of 34% to 41% GHG emission per year per household. On the other hand, bike sharing has the potential to provide a transportation free of emissions (Shaheen et al., 2010a). But the environmental costs have to be considered in the balancing and relocation process (Pal and Zhang, 2017). Bike sharing has helped to reduce CO₂ in some cases from 47.000 to 150.000 kg per year. A considerable reduction has been shown in Hangzhou with a decrease of 191.000 kilograms CO₂ per day (Shaheen et al., 2012).

Congestion reduction and transport infrastructure efficiency . The effect of car own-

ership reduction has an prominent effect of the reduction of the vehicles found in the transportation system. [Shaheen and Chan \(2015\)](#) found that depending on the study location, each car sharing vehicles replaces 4 to 13 private cars. In North America until 2010, [Martin et al. \(2010\)](#) estimated a reduction of 90.000 to 130.000 cars because of car sharing. This leads to a more efficient use of the existing road infrastructure and fewer space ([Shaheen and Cohen, 2012](#)).

Increment of transit share When a car sharing user does not have an own vehicle, the probabilities are higher to chose other means of transport such as public transport, walking and cycling ([Kent, 2014](#)). Shared mobility serves to connect the first or last link to public transport. Therefore, it increases the connectivity the "first and last mile," enabling the access to public transportation to places where private cars where the only possible mode (([Ricci, 2015](#)), ([Shaheen and Cohen, 2012](#))). Consequently, shared mobility increases public transport use. On a survey in North America, more than the 95% of the respondents agreed that bike sharing enhance public transport and 41% agreed that they had made trips on public transportation or bike sharing instead of a private car ([Shaheen et al., 2012](#)).

Increment of active modes share The first year of bike sharing implementation there was a 44% increase bicycles usage in Lyon ([Shaheen et al., 2010a](#)). Furthermore, after bike sharing systems, modal shared of bicycle increased in Barcelona and 1.5% in Paris. In conclusion, bike sharing might encourage cycling to users than in other cases would not use this transport mode ([Shaheen et al., 2012](#)). Finally, 64% respondents of a survey agreed that they walk more after joining a bike sharing program ([Shaheen et al., 2012](#)).

Health benefits Change of travel behavior from exclusively private car usage to multimodal travel behavior can be considered to have health benefits. Shared mobility participants are likely to interact with others in the neighborhood, increasing the sense of inclusion and belonging. This fact might decrease mental health problems [Kent \(2014\)](#). Moreover, cycling and walking as exercise has a good positive effect on health ([Büttner and Petersen, 2011](#)).

Individual economic savings Shared vehicles are more flexible and less expensive than owning one. Users of shared vehicles can get the benefits of private vehicles without fixed costs such as purchase costs, insurance, storage, and maintenance (([Shaheen et al., 2012](#)), [Shaheen and Cohen \(2012\)](#)), leading to individual financial savings ([Cheng, 2016](#)). Additionally, the reduction of VMT decreases driving costs ([Shaheen et al., 2012](#)). In conclusion, car sharing might represent a monthly household saving from 154 to 435 American dollars in the USA after becoming a member ([Shaheen and Chan, 2015](#)).

Less consumption of raw materials If the car ownership decreases, less raw materials are required for private cars construction ([Firnorn and Müller, 2011](#)).

Increase equity of access Private cars are related to an access inequality to work, education, services where they are only accessible by these transport mean. Vehicles sharing can moderate this unequal access ([Kent, 2014](#)). Household users who require periodic vehicle access can have the benefits of private vehicles without paying the full costs of car ownership and the investment to buy one. This cost reduction is also beneficial for college students and low-income households ([Shaheen et al., 2012](#)). However, [Tyndall \(2017\)](#) showed that the users of car sharing in the USA are usually white, young, educated and employed.

Higher environmental awareness As a consequence of the ecological benefits, shared mobility users have reported a higher degree of environmental awareness after joining a shared mobility program ([Shaheen et al., 2012](#)).

Enhancement of the image of cities and support tourism Shared mobility enhance the

image of a city. It shows that a city cares about sustainable transport and environmental awareness. Also, tourists can visit the city by bicycle saving money for several rides on public transport (Büttner and Petersen, 2011).

2.2 Model building and selection

To estimate or predict relationships among variables and to select the most influencing of them, a model can be built. A model is "a simplification of the reality," so there is never going to be the perfect model (Posada and Buckley, 2004). This section provides an overview of the models that can be applied for the purposes of this thesis. They are selected based on their simplicity of implementation and based on the particular aims of this study, i.e. the exploration of the factors that affect the deployment of shared mobility systems. Specifically, regression models are discussed and particular three types: multiple linear regression models (Section 2.2.2), generalized linear models (Section 2.2.3) and regression trees (Section 2.2.7, Section 2.2.8). In the next sections, the theoretical background of the models is presented including aspects of modeling that are relevant for this study, such as Model selection, assessment and validation (Section 2.2.1), model diagnostics (Section 2.2.4), aspects of collinearity (Section 2.2.5).

2.2.1 Model selection, assessment and validation

The best model is commonly based on finding the equilibrium between bias and variance by changing the number of parameters in the model (Posada and Buckley, 2004). If the number of parameters decreases, the variance decreases, however the bias increases and vice versa. In other words, we want to account for the variance as much as possible, which means as many variables as possible but on the other hand, we want to keep the model simple with as few variables as possible. There are two different approaches to find the best model. 1) *Model*

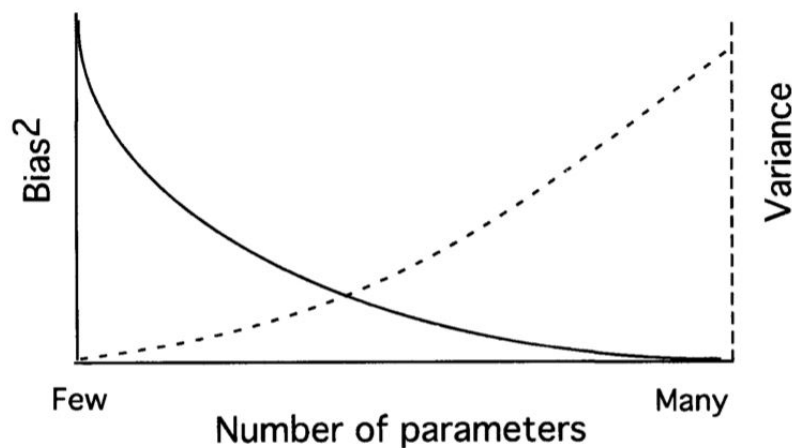


Figure 2.4: The number of parameters
Source: Posada and Buckley (2004)

selection that estimates the performance of different models and choose the best one. 2) *Model assessment* that estimates the error of the chosen final model (James et al., 2013). The widely used approach is to split the data set into three groups:

- a) Training set (50% of the data): Used to build the models
- b) Test set (25% of the data): Used to estimate the error of the models to select the best one
- c) Validation set (25% of the data): Used to calculate the error of the final chosen model.

To analyze the accuracy of the model, i.e., how well the model fit the data, there are two measures that are commonly used: mean square errors (MSE) and the R^2 value.

Let y_i be an observation and \hat{y} be the fitted value, then the MSE is roughly the average amount of \hat{y} that will deviate from the original observation y_i .

$$MSE = \frac{SSE}{n} \quad (2.1)$$

Where n is the number of observations, and SSE is the sum the squared errors. SSE is the distance from each the point to the fitted model.

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.2)$$

where, y_i is the observation or response and \hat{y}_i is the predicted value citejames2013introduction.. The R^2 value measures the relationship between the Y observations and the \hat{Y} predictions.

$$Cor(Y, \hat{Y}) = R^2 = \left(\frac{\sum (y_i - \bar{y})(\bar{y}_i - \bar{\hat{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (y_i - \hat{y})^2}} \right)^2 \quad (2.3)$$

where, \hat{y} is the fitted observations, \bar{y} is the mean of the observations, and $\bar{\hat{y}}$ is the mean of the fitted observations. R^2 is close to one when the data fit well to the model. However, SSE and R^2 are not suitable to select the best model with different number of parameters. Thus, we can additionally estimate (in a direct or indirect way) the error (James et al., 2013).

1. Indirect model selection methods

The most used indirectly methods are the adjusted R^2 , Akaike Information Criterion (AIC), or Bayes Information Criterion (BIC) (Chatterjee and Hadi, 2015).

- (a) **Adjusted R^2** (R_{adj}^2) is used to adjust the unequal number of variables of different models that R^2 does not consider.

$$R_{adj}^2 = 1 - \frac{n-1}{n-k-1} (1 - R^2) \quad (2.4)$$

where, $n-k-1$ is the number of degrees of freedom and SSE is the sum the squared residuals. A higher number of k parameters penalizes the error. If the ratio n to k is large than the regression model tends to have lower variance.

- (b) **The Akaike Information Criterion (AIC)** (Akaike, 1973) is also used as a models' selection criteria. AIC has the advantage that helps to compare two nested models. Models are preferred if they have a lower AIC. For two models with the same SSE, AIC penalizes the one with more parameters. However, it has the disadvantage that it tends to improve with a larger number of k parameters, thus it is commonly accused of being prone to allow overfitted models to be selected.

$$AIC = n * \ln(MSE) + 2k \quad (2.5)$$

- (c) **Bayesian Information Criterion (BIC)** (Schwarz, 1978) tends to control the overfitting of AIC. BIC is proportional to AIC but instead of using a factor of 2, it uses $\log(n)$. Then, BIC tends to select simpler models assuming that $\log(n) > 2$ ($n = e^2 = 7.4$).

$$BIC = n * \ln(MSE) + k * \ln(n) \quad (2.6)$$

2. Direct model selection methods

Direct methods allow to check directly the errors for each model and therefore, select the model with the smallest estimated residuals. Another application is the control of overfitting, which is a misinterpretation of results that appear as significant but cannot reproduce the model for a new sample. To identify this problem train/test split can be realized using either cross-validation or the validation approach presented below.

- (a) **Validation approach** starts by dividing the data set into training set and test set. Then, the training set is used to build the model and the test set to predict the responses. The residuals of the fitted and predicted values are usually measured by MSE. The process continues by selecting a new training set and test set in order. The disadvantage of this method is that the MSE can be very variable for each sampling set and by using a training set that has a lower number of observations might overestimate the error (James et al., 2013).
- (b) **k-Fold Cross-Validation** deal with the drawbacks of the validation approach. It divides the data set into equal-sized k subsets or *folds*. The first fold is the test set, and the $k - 1$ sets are used for fitting the model. Then, the MSE_1 is calculated for the k th fold that was left out. We repeat this procedure by taking the next k fold as a test set and calculation MSE_2 until we calculate the MSE_k . Then the final k -fold CV is the average the errors:

$$CV_k = \frac{1}{k} \sum_{i=1}^k MSE_i \quad (2.7)$$

Common k values are from 5 to 10; higher values would have an high computational cost. Another advantage of these values is the bias-variance trade-off, 5 to 10 folds empirically have shown neither a high bias nor a very high variance.(James et al., 2013).

2.2.2 Multiple Linear Regression

To build a multiple linear regression, we denote that a vector Y with size $n \times 1$ is linearly related to a $n \times (k + 1)$ matrix X of k parameters and presenting ϵ residuals.

$$Y = X\beta + \epsilon \quad (2.8)$$

where, β is nonzero $(k + 1) \times 1$ vector of coefficients (Miller, 1984). Each observation y_i can be also written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i, \quad \text{for } i = 1, 2, \dots, n \quad (2.9)$$

Usually, β_0 is referred as the intercept constant. We estimate the parameters $\beta_0, \beta_1, \dots, \beta_k$ by using the ordinary least squares (OLS) method. OLS minimizes the sum of the square errors (SSE) (distance from each the point to the line), where the predicted values \hat{y}_i use the estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ (Chatterjee and Hadi, 2015).

For the estimated values the equation becomes:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_k x_{ik}, \quad \text{for } i = 1, 2, \dots, n \quad (2.10)$$

2.2.3 Generalized Linear Models (GLM)

Each observation y_i can be also expressed as:

$$y_i = \mu_i + \epsilon_i \quad (2.11)$$

The Generalized linear models (GLM) are an extension of OLS and they are used when OLS are not appropriate. Usually, this family of models estimation is performed by the method of maximum likelihood. GLM assume that the error ε_i presents a distribution from the exponential family, such as: binomial, Poisson, Gaussian. Also, they consider the mean function μ_i as a function of the linear observations:

$$h(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} \quad (2.12)$$

where, $h(\mu_i)$ is a function that links μ_i with the observation Y_i . The function has to be monotonic and differentiable. As an example Poisson regression models use the logarithmic function as link and the random component has a Poisson distribution. These are appropriate for count data. Another examples are to use logistic regression that use a binomial distribution and a logit function as link or a gaussian distribution with a linear function link (Chatterjee and Hadi, 2015).

2.2.4 Model Diagnostics

The main potential problems in linear models are:

1. *Non-linearity of response-predictor relationship.* Linear models have the assumption that they have a linear relationship between the predictors and the response. Residuals plots are helpful to diagnostic this problem. They are scatter-plot of the error vs the predicted response. If these plot have a non-linear association, we assume there is a non-linearity of response-predictor relationship. A possible solution is to use non-linear transformation on the predictors on the regression, such as $\log(X)$, \sqrt{X} , X^2 (James et al., 2013).
2. *Heteroscedasticity* Linear models assume that the error have a constant variance. However, heteroscedasticity is when non-constant variance of error terms are present in the model. The variance error might increase with the fitted value, i.e., in the residual plot might present a "funnel shape". A possible solution is to transform the observations Y to a concave function, such as $\log(Y)$, or \sqrt{Y} (James et al., 2013). For instance, the model for $\log(Y)$ will be:

$$Y = e^{X*\beta} \quad (2.13)$$

Another transformation is the so called boxcox (Box and Cox, 1964):

$$\mathbf{z}_{\lambda,i} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda * \hat{y}^{\lambda-1}} & \text{if } \lambda \neq 0 \\ \lambda * \ln(\mathbf{y}_i) & \text{if } \lambda = 0 \end{cases} \quad (2.14)$$

where, \hat{y} is the geometric mean of the observations: $\hat{y} = (y_1 * y_2 * \dots * y_n)^{1/n}$ and λ is a parameter. Then the goal will be to have the model

$$\mathbf{z}_\lambda = X\beta + \epsilon \quad (2.15)$$

that minimizes the value of λ . In order to find the best value for λ , we use a range of values, usually from -2 to 2 with intervals of 0.1. Then, we calculate the square sum of the regression (SSR) for each λ . Finally we choose the λ value that results on the maximum SSR value. For previously mentioned transformations, the whole observations have to be positive. Therefore, we can add a constant value to all the observation for not having zero or negative values (XXXXXXXXXX models lineages aplicados).

3. *Outliers* An outlier is an observed value that is far from the predicted value about the other observations and their predicted values. Fitted values with a studentized residual

greater than 3 have higher possibilities to be an outlier. A studentized residual is the result of the division of a residual e_i divided by the estimated error. If the outlier is produced because a mistake on the data collection the best solution is not to consider the outlier in the model (James et al., 2013).

4. *High leverage points* High leverage points are observations where the predictors x_i present an unusual value. However, in large data sets, it can be that some predictors are out of range.
5. *Collinearity*. See section 2.2.5.

2.2.5 Collinearity

We assume that all the parameters are independent from each other. However, some parameters can be correlated with each other and might change the regression coefficients (tends to inflate the variance of at least one coefficient) of the model or produce larger standard errors and tend to be unstable (Mack, 2016). In OLS, collinear variables the effect of the collinear variables is not separated (James et al., 2013).

To deal with this issue, first, we have to detect the collinearity, and then we correct it.

- I) **Detecting collinearity.** A common way of detecting collinearity is a correlation matrix, containing correlation coefficients for each pair of variables. The diagonal elements of this matrix are equal to one. Thus, if a coefficient is close to one, it means that the parameters are highly correlated. If two parameters are completely uncorrelated, we called them orthogonal (Chatterjee and Hadi, 2015). The two most common methods to estimate correlation coefficients are Pearson's correlation coefficient and Spearman's rank correlation coefficient (Hauke and Kossowski, 2011). Pearson's correlation coefficient measures linearity relationship between two variables. It assumes the normality of the variables analyzed. Spearman's rank correlation coefficient is a nonparametric rank statistic. It calculates the monotonic association between two variables. It does not assume that the variables behave linearly. After a literature review, Bishara and Hittner (2012) concluded that Spearman correlation would be more valid than Pearson's test for nonnormal variables by applying a nonlinear transformation and then performing a Pearson's correlation on the transformed data.

There are mainly three criteria to detect collinearity:

- (a) If the eigenvalues λ present a uniform value, there is not collinearity. Each correlation matrix A of k parameters have k eigenvalues $\lambda : \lambda_1, \lambda_2, \dots, \lambda_k$, where $|A - \lambda I| = 0$. So we use the condition number κ :

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}} > \theta \quad (2.16)$$

θ in approximately 10^2 for (Mack, 2016) and 15^2 for (Chatterjee and Hadi, 2015)

- (b) Empirical criterion:

$$\sum_{j=1}^k \frac{1}{\lambda_j} > 5k \quad (2.17)$$

- (c) The variance inflation factor VIF_j measures the relationships between the predictor variables. If VIF_j is greater than 10 we assume collinearity.

$$VIF_j = \frac{1}{1 - R_j^2} > 10 \quad \text{for} \quad j = 1, 2, 3, \dots, k \quad (2.18)$$

- II) **Addressing collinearity.** According [Chatterjee and Hadi \(2015\)](#), simply deleting the collinear parameters does not always solve the problem. However, the most common approach in practice is to remove the most correlated variables. An alternative to address this problem is the principal component analysis (PCA) and ridge's regression.

Principal component CA The principal components method is based on changing the original parameters into orthogonal variables. Each principal component C_1, C_2, \dots, C_k is a linear function of the standardized variables $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_k$. The standardized variables have mean equals to zero and standard deviation equal to one.

$$C_j = v_{1j}\tilde{X}_1 + v_{2j}\tilde{X}_2 + \dots + v_{kj}\tilde{X}_k, \quad j = 1, 2, \dots, k \quad (2.19)$$

$$C = \tilde{X}V \quad (2.20)$$

The coefficients v_j are the j th component of the j th eigenvector corresponding to λ_j . The *eigenvector* (v) is defined as:

$$Av = \lambda v \quad (2.21)$$

The PC's are arranged in a way that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$. Another property is that the variance of C_j is equal to λ_j . So the PC with the largest variance is the first one. Finally, we can create a regression with the matrix C instead of the matrix Z :

$$Y = C\alpha + \epsilon' \quad (2.22)$$

Then, to find the initial regression coefficients β , we can use the expression:

$$\beta = V\alpha \quad (2.23)$$

There is not a dimensional reduction if we use all the components in the regression. So they have to be selected to perform this effect. However, after a principal components regression is implemented, the resulting model is difficult to interpret because it does not realizes a variable selection, including all the *parameters* in the final model ([James et al., 2013](#)).

2.2.6 Variables Selection

When the number of observation n is large, we want to construct an equation with a subset of parameters. A subset or reduced model is a model with some missing terms ([Mack, 2016](#)). To change the number of parameters, the problem is which parameters should be included and if the parameters should be included as the original X , or X^2 or $\log(X)$. Usually, first, we decide which variables are going to be taken into account, and then we investigate the form that they should be included. Finally, there is not the "best subset" of parameters, but there are several that might be adequate. Another reason to reduce the number of possible predictors is the lower cost to calculate them, the prediction will be more accurate by eliminating variables that do not provide information, and the regression coefficients will present smaller standard errors ([Chatterjee and Hadi, 2015](#)).

The technique called **F-Test** analyzes if the reduced model (RM) is adequate in relation with the (FM), the null hypothesis is that RM is adequate against H1, that states that FM is adequate. Thus, to see if the reduced model is optimal we use the F - Test (2.24) where p is the number of parameters selected for the RM. So H_0 is rejected if $p(F) \leq \alpha$ ([Chatterjee and Hadi, 2015](#))

$$F = \frac{[SSE(RM) - SSE(FM)]/(k + 1 - p)}{SSE(FM)/(n - k - 1)} \quad (2.24)$$

Nevertheless, with F -Test we cannot compare models with different parameters (nested models).

The simplest method is to erase the variables. However, this can decrease the variances of the estimates. These variables have a nonzero regression coefficient, so the coefficients of the remained variables may have a lower variance in the RM, and higher bias and thus, loss of precision in the prediction. Ordinary Least-squares methods are usually implemented to find best-fitting subsets of parameters and for control. The goal is to have the closest fit of the selected parameters with a set of observations Y by minimizing the MSE of the prediction (Chatterjee and Hadi, 2015). Two common variable selection methods are 1) stepwise regression and 2) the lasso technique.

1. **Stepwise regression** This method helps to address the best subset selection for a large number of k parameters. There are three types of stepwise selection procedures: 1) forward selection, 2) backward selection 3) stepwise selection (both directions). These procedures are useful for **noncollinear** parameters.

- (a) **Forward selection.** For the forward selection, the equation starts only with the constant term (i.e., no parameters). Then, the first parameter is the one with the highest correlation with the observations Y . If the regression coefficient is significantly different than zero then the parameter is retained and the second variable with the second highest correlation is selected, then the regression coefficient is calculated and if it is significant the parameter stays. Then the third variable is tested in the same way, and the process continues until the last variable. The significance of the coefficients is judged by a t -Test.
- (b) **Backward selection.** The backward elimination process, on the other hand, starts with a full equation. Then, the variables are eliminated according to their contribution to the reduction of the SSE. The first deleted variable is the one the contributes the less to reduce the SSE. This is similar as deleting the variable with the least t -Test. The process finishes when the whole non-significant variables are removed. The backward selection is the best to handle collinearity.
- (c) **Both directions.** The stepwise method in both directions is one of the most common techniques. It consists on sequentially add or delete parameters. It starts with a forward selection, but at each step, it has the possibility to delete a parameter. The advantage is if a nonsignificant variable enters in the process, it might be eliminated later.

AIC and BIC might be used as a criterion to add or eliminate the parameters. The process finishes when the addition or reduction of parameters do not reduce significantly the AIC or BIC. The information criteria with the t -Test differ in a significant way. t -Test calculates the significance of the variables and AIC (BIC) just the reduction of the criterion (Chatterjee and Hadi, 2015). Among other methods for variable selection, stepwise regression has shown high accuracy but low computational efficiency (Lin et al., 2011).

2. **Lasso technique** The least absolute shrinkage and selection operator (lasso) technique (Tibshirani, 1996) shrinks the coefficients β increasing stability and while retaining the bests variables. In other words, lasso does variables selection and shrinkage. A significantly reduce their variance can be preformed after shrinking the coefficients. Lasso assumes that X_{ij} are standardized with a mean of zero and a standard deviation of 1. And the observations Y_i are centered. Lasso minimizes the sum of the squared differences between the observation and the linear regression. However, it is restricted for the sum of the absolute values of the coefficients to be less than a positive turning parameter t .

$$\hat{\beta}_{lasso} = \operatorname{argmin}_{\beta} \sum_{i=1}^n (Y_i - \beta_0 - \sum_{j=1}^k X_{ij}\beta_j)^2 \quad (2.25)$$

Subject to:

$$\sum_{j=1}^k |\beta_j| < t, \quad t \geq 0 \quad (2.26)$$

Let $t_0 = \sum_{j=1}^k |\beta_j|$, $t < t_0$ will shrinkage the β coefficients to zero. The equivalent form of lasso is:

$$\hat{\beta}_{lasso} = \operatorname{argmin}_{\beta} \left(\sum_{i=1}^n (Y_i - \beta_0 - \sum_{j=1}^k X_{ij} \beta_j)^2 + \lambda \sum_{j=1}^k |\beta_j| \right) \quad (2.27)$$

By including the absolute value in the penalty $\lambda \sum_{j=1}^k |\beta_j|$ (also called l_1 penalty), we assure that the coefficients will shrink to zero, and some of them might also be exactly zero producing an exclusion of the variables. Models generated with lasso are easy to interpret because they preform a subset of the variables. The selection of the λ value is critical in this technique because depending on this values are the number of variables considered for the model. Therefore, we have to calculate the cross-validation error for each value of λ and select the λ for the smallest error. However, to have a simpler model, with fewer variables, we can choose the "largest value of λ such that error is within one standard error of the minimum" (James et al., 2013). Empirically, stepwise regression has shown more accuracy but less computational efficiency than lasso technique (Lin et al., 2011).

2.2.7 Decision tree methods

Decision tree or regression trees methods are used for classification and regression. They involve a segmentation if the predictor's data set into simple regions. For the observations, the mode or mean of each region are used. The set of rules can be summarized in a tree. The method involves the combination of several trees. More trees lead to a harder interpretation but also to a higher accuracy (James et al., 2013). The process consists to divide the parameters $x_1, x_2, x_3, \dots, x_k$ into J no overlapping regions: $R_1, R_2, R_3, \dots, R_J$. Then, each region R_j will contain the mean of all the observations that belong to the region. For simplicity, the regions R_j are rectangles; therefore, they are also called boxes. Then, the goal is to minimize the number of boxes to have the lower SSE.

$$\sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \quad (2.28)$$

Where \hat{y}_{R_j} is the mean of the observation inside the region R_j . Since this optimization is computationally unfeasible, we consider the recursive binary splitting, which is a top-down greedy approach. It begins from the top of the tree with all the observations in a single region. Then, it is split into two new branches. It does not looks ahead for the splitting but taking the best split in each step that will lead to the better tree in the further step. So we chose a parameter x_j and a cut-point s that leads to the lowest reduction of the SSE. We do this for each parameter $x_1, x_2, x_3, \dots, x_k$ to find the value j and s that minimizes:

$$\sum_{i x_i \in R_{1(j,s)}} (y_i - \hat{y}_{R_1})^2 + \sum_{i x_i \in R_{2(j,s)}} (y_i - \hat{y}_{R_2})^2 \quad (2.29)$$

where, \hat{y}_{R_1} is the mean of the observations where $x_j < s$ and \hat{y}_{R_2} is the mean of the observations where $x_j \geq s$. This process is repeated until a stopping criterion is reached, creating $R_1, R_2, R_3, \dots, R_J$ boxes. The criterion, for instance, can be to build the trees until the regions do not have more than a minimum number of observations (James et al., 2013). However, this process may lead to overfitting, since the resulting tree might have high complexity. A smaller tree with fewer regions $R_1, R_2, R_3, \dots, R_J$ might lead to a result with less variance but a better

interpretation. To achieve this, a tree must be so long until the change of the reduction of the SSE is lower than a threshold. Nevertheless, this approach might not lead to the best split, so a better strategy is to start with a very long tree T_o and "prune" it into smaller subtrees T . Then the goal would be to select the best subtree through cross-validation or validation approach. To select a small set of subtrees we use the cost complexity pruning. It considers a non-negative turning parameter α that indexes a sequence of trees minimizing:

$$\sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T| \quad (2.30)$$

where, $|T|$ is the number of nodes in the tree T . R_m is the box of the m th node, and \hat{y}_{R_m} is the mean observation of the box R_m . α control the trade-off between the fit of the subtree with the data and its complexity. If $\alpha = 0$, then $T = T_o$; while α increases then the subtrees will be smaller. α can be selected after a cross-validation approach (James et al., 2013).

2.2.8 Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) is machine learning algorithm that performs regression, classification and ranking (Friedman, 2001). It is a mix of gradient descent and boosting. If the data have a linear distribution, linear models are likely to perform better, but for nonlinear relationships, decision trees might fit the data in a better way. However, trees usually are less accurate than other regression approaches.

The boosting approach helps decision trees to improve their prediction. Boosting is a procedure to reduce the variance of a model. It involves the creation of multiple B training sets. Then, it builds a prediction for each training set $\hat{f}^1(x), \hat{f}^2(x), \dots, \hat{f}^B(x)$ and it fits different decision trees to each copy. Each tree is a modified version of the original data set, and they grow sequentially by using the information of the previously grown tree. Therefore, these trees are dependent on each other. The residuals are fit to the decision tree, rather than a single decision tree to the data. We choose the sample data that modeled poorly in the system before, i.e., in areas where the system is not performing well. Then, the residuals are updated after adding the new decision tree into the fitted function. Finally, it combines all the trees to create a single model. Algorithm 1 summarizes the boosting approach:

Algorithm 1 Boosting for regression trees

- 1: $\hat{f}(x) \leftarrow 0$
 - 2: $n \rightarrow$ observations in training set
 - 3: **for** $i = 1 \rightarrow n$ **do**
 - 4: $r_i = y_i$
 - 5: **for** $b = 1 \rightarrow B$ **do**
 - 6: Fit a tree's prediction $\hat{f}^b(x)$ with d splits ($d + 1$ nodes) to training data (X, r)
 - 7: Update $\hat{f}(x)$ by adding a shrunken new tree: $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda_{gbm} \hat{f}^b(x)$
 - 8: Update the residuals: $r_i \leftarrow r_i + \lambda_{gbm} \hat{f}^b(x)$
 - 9: *output model:*
 - 10: $\hat{f}(x) = \sum_{b=1}^B \hat{f}^b(x)$
-

Source: James et al. (2013)

The number of nodes of the trees is determined by the number of splits d , while $d = 1$ usually works well. An extra shrinkage parameter λ_{gbm} controls the "learning" rate. λ_{gbm} is a small positive number which usually takes the numbers of 0.01 or 0.001. High values of B require

smaller λ_{gbm} . Boosting can overfit after a big B . Therefore, a cross-validation approach is useful to select B (James et al., 2013). 5 to 10-fold cross validation are recommended (Ridgeway, 2006).

The resulting model might be difficult to interpret. Therefore, variable importance plots can be constructed from a relative importance measure for each predictor. Variable importance describes how much predictor contributes to the fitted model. We can obtain the relative contributions by averaging over all B trees the total amount that the SSE is reduced after the splitting over a given predictor". An important predictor would have a higher value. A faster approximation to find the model is to consider a differentiable loss criterion that can be derived by a numerical optimization. The loss of using $\hat{f}(x)$ to predict y is:

$$L(f) = \sum_{i=1}^n L(y_i, \hat{f}(x_i)) \quad (2.31)$$

The goal would be to "minimize $L(f)$ with respect to f , where $f(x)$ is constrained to be a sum of trees".

$$\hat{\mathbf{f}} = \underset{\mathbf{f}}{\operatorname{argmin}} L(\mathbf{f}) \quad (2.32)$$

where, \mathbf{f} ($\mathbf{f} = \hat{f}(x_1), \hat{f}(x_2), \dots, \hat{f}(x_n)$) are the values from the approximation function ($\hat{f}(x_i)$). This involves often to express the solution in the form:

$$\mathbf{f}_B = \sum_{b=0}^B \mathbf{h}_b \quad (2.33)$$

where, $\mathbf{f}_0 = h_0$, each \mathbf{f}_b is based on the current vector \mathbf{f}_{b-1} . Numerical optimization methods differ of computing each step \mathbf{h}_b . Steepest descent is one of the most used and simple minimization methods. It defines $\mathbf{h}_b = -\rho_b \mathbf{g}_b$, where ρ_b is a scalar and \mathbf{g}_b is the gradient of $L(\mathbf{f})$. Therefore, at each step these parameters have to be calculated:

$$g_{ib} = \left[\frac{\partial L(y_i, \hat{f}(x_i))}{\partial \hat{f}(x_i)} \right]_{\hat{f}(x_i) = \hat{f}_{b-1}(x_i)} \quad (2.34)$$

$$\rho_b = \underset{\rho}{\operatorname{argmin}} L(\mathbf{f}_{b-1} - \rho \mathbf{g}_b) \quad (2.35)$$

After the estimation, the current solution is updated: $\mathbf{f}_b = \mathbf{f}_{b-1} - \rho \mathbf{g}_b$. It should be mentioned that regarding the loss function, a Gaussian function would be adequate for minimizing squared error and the Laplace for minimizing the absolute error (Friedman et al., 2001).

Algorithm 2 shows the merge of the gradient descent with boosting:

If the subsample P is preformed, it is a number higher than zero. A value of 0.5 is recommended (Ridgeway, 2006). It evaluates the performance of the prediction. Usually, six terminal nodes K do an excellent job (James et al., 2013) and 3,000 to 10,000 number of trees are considered within a shrinkage range from 0.01 to 0.001

2.3 Factors affecting the deployment of car and bike sharing

The benefits of shared mobility are dependent on exogenous and endogenous factors that might influence the deployment of a shared mobility system (Büttner and Petersen, 2011).

– **Endogenous factors** – include those that can be adjusted according the external factors of the system. There are two types of endogenous factors:

- *Physical design*: hardware and technology (access technology, vehicles, stations, software) and service design (size, density, registration, information, target groups, availability, charges, public transport integration).
- *Institutional design*: operators, contracts, costs and financing.

Algorithm 2 *Gradient Tree Boosting*

```

1:  $L \rightarrow$  a loss function
2:  $B \rightarrow$  the number of iterations
3:  $K \rightarrow$  terminal nodes of each tree
4:  $\lambda_{gbm} \rightarrow$  the shrinkage parameter
5:  $P \rightarrow$  subsampling rate
6:  $\hat{f}_0(x) = \operatorname{argmin}_{\rho} \sum_{i=1}^n L(y_i, \rho)$ 
7:  $n \rightarrow$  observations in training set
8: for  $b = 1 \rightarrow B$  do
9:   for  $i = 1 \rightarrow n$  do
10:     Compute the negative gradient:  $z_{ib} = \left[ \frac{\partial L(y_i, \hat{f}(x_i))}{\partial \hat{f}(x_i)} \right]_{\hat{f}(x_i) = \hat{f}_{b-1}(x_i)}$ 
11:     Select randomly select  $P * n$  cases from the data set
12:     Fit a regression tree with  $K$  terminal nodes with the randomly selected observations
        to target  $z_{ib}$ 
13:     for  $k = 1 \rightarrow K$  do
14:        $\rho_{kb} = \operatorname{argmin}_{\rho} \sum_{x_i \in S_k} L(y_i, \hat{f}(x_i) + \rho)$  where,  $S_k$  is the set of parameter within
        terminal node  $k$ 
15:       Update  $\hat{f}(x)$ :  $\hat{f}_b(x) \leftarrow \hat{f}_{b-1}(x) + \sum_{k=1}^K \lambda_{gbm} \rho_{kb} I$ 
16: output model:
17:  $\hat{f}(x) = \hat{f}_B(x)$ 

```

Source: [Friedman et al. \(2001\)](#), [Ridgeway \(2006\)](#)

– **Exogenous factors** – are those independent of shared mobility systems. They correspond to specific features and characteristics of a city that difficult to change, such as:

- *City size*: large (>500,000 inhabitants), medium (100,000 to 500,000 inhabitants), small (20,000-100,000 inhabitants).
- *Weather*: Table 2.1 indicates the studies of shared mobility that took into account the weather. Usually for car sharing they consider temperature and precipitation, but for bike sharing they include other weather variables such as wind speed, relative humidity and mean hours of light.
- Mobility behavior,
- Population and jobs density,
- Demographic factors,
- Economic factors,
- Geographic factors and topography: hilliness,
- Existing infrastructure,
- Financial situation,
- Political situation,

Table 2.2 shows some examples of factors analyzed in shared mobility researches.

– **Time** – Time is also a factor that influence on arrivals and departures of shared mobility systems. Table 2.3 shows some studies that considered time clusters in the bookings of car and bike sharing. The majority classify the day type, and they also subdivide the day into time intervals.

Table 2.1: Weather variables in shared mobility studies

System	Author	Temperature	Precipitation	Relative humidity	Wind speed	Mean hours of sunlight
SBBS	1	✓	✓	✓	✓	
SBBS	2	✓	✓	✓		
SBBS	3	✓	✓			✓
FFCS	4	✓	✓			
FFCS	5	✓	✓			

1=Chardon et al. (2017) 2=Faghih-Imani et al. (2014) 3=Caulfield et al. (2017) 4=Schmöller et al. (2015) 5=Schmöller and Bogenberger (2014)

Table 2.2: Studied exogenous factors in a literature selection

Variables	A	B	C	D	E	F	G	H	I
Population density	✓	✓	✓	✓	✓	✓		✓	
Job density		✓				✓			
Neighborhood age				✓					
Education level share	✓			✓					
Vehicle ownership rate	✓			✓					
Inhabitants age share		✓	✓					✓	
Household size share		✓	✓						
Rent prices		✓							
Land use type share		✓	✓						
Registered vehicles			✓						
Bus lines serving the district			✓						
Mode to commute	✓								
Public transport station					✓				✓
Bike sharing station (dist.)	✓				✓				
CBD (dist.)						✓			
Railways (dist.)					✓				
Services (dist.)									✓
Restaurants, Coffee (dist.)									✓
Commercial enterprises (#)		✓			✓				
Roads length					✓				
Cycling ways length	✓								
Number of stations of BS	✓								
On-street parking capacity				✓					
Walkability				✓					
PT stations (#)	✓								
Frequency of use			✓				✓		
Use of the stations (daily)							✓		
Traveled distance							✓		
Vehicle Availability (CS)					✓				
Docks per station (BS)	✓								
Frequency OD pair							✓		

A=Chardon et al. (2017), B=Schmöller et al. (2015), C=Kang et al. (2016), D=Celsor and Millard-Ball (2007), E=Comendador et al. (2014), F=Faghih-Imani et al. (2014), G=Caulfield et al. (2017), H=Schmöller and Bogenberger (2014), I=Willing et al. (2017)

There is a lack of studies that treat bike and car sharing together. However, Efthymiou et al. (2013) have studied individual factors from a survey of young people (18-35 years old) that might affect bike and car sharing systems, such as environmental consciousness, household

Table 2.3: Time clustering types

Vehicle type	Author	Day type	Day time interval
SBBS	Faghih-Imani et al. (2014)	Workday, weekend	AM, midday, PM, Friday and Saturday night
SBBS	Caulfield et al. (2017)	Workday, weekend	AM peak, PM peak, AM off peak, PM off peak
FFBS	Reiss and Bogenberger (2017)	Workday, weekend	According to demand (0-6,6-10,10-16,16-20,20-24)
FFCS	Schmöller et al. (2015)	Day of the week, official holiday	0-6am, 3 hours intervals
FFCS	Willing et al. (2017)	Workday, weekend	4 hours interval

income, and size, vehicles kilometer traveled per week, transport mode to different trips, and time to work/school. Clark and Curl (2016) studied the social inclusion of car and bike sharing. However, there are just few that study together car and bike sharing. In the literature, there a growing number of studies that have explored the exogenous factors that affect shared vehicles. For example, Eftymiou et al. (2013) examined individual factors that might influence the adoption of bike and car sharing systems by surveying young people (18-35 years old) in Greece by implementing ordered logit models. The factors included, for example, environmental consciousness, household income and size, vehicle kilometers traveled per week, transport mode to different trips, and travel time to work or school. However, this study is one of a few that deal car and bike sharing together.

2.3.1 Factors affecting the deployment of car sharing systems

Regarding station based car sharing (SBCS), Kang et al. (2016) investigated the factors affecting the demand in Seoul. The independent variables such as the built environment, demographics, and transportation attributes, were measured for each city district. After a linear regression model, the variables affecting the bookings' intensity are the size of the business area, the share of inhabitants between 20 and 39 years old, the total number of registered cars, and subway entrances.

Celsor and Millard-Ball (2007) studied the market potential of SBCS using a supply model based on 13 regions in the USA to determine where the demand is concentrated. The two market segments were the demographic groups more likely to join the SBCS system and the geographic areas (neighborhoods) where car sharing might have a better performance. The relationship between supply and neighborhood characteristics was demonstrated after a Pearson's correlation analysis between the level of service, i.e., the number of shared vehicles in a neighborhood in a half mile radius, and 16 independent variables, like demographics, commute mode share, vehicle ownership, and neighborhood characteristics. Most variables presented a high degree of correlation, such as density, one-person households, transit and walking share; however, some variables were not correlated, like income, education and bicycle commute.

Kortum et al. (2016) discussed the growth rates and success factors of FFCS in 34 cities in nine countries (mainly Germany and the US). In these cities, there is a generalized growing trend in the number of bookings per day as the car sharing programs continue operating. Some cities as Berlin and Munich showed that the growth had been slowed, probably because they

are close to the saturation point. The two main success factors were the household size and the inhabitant's density.

Willing et al. (2017) correlated the densities of the rentals of a FFCS system with the densities of the points of interest in the city of Amsterdam. They implemented for the regression a generalized linear model with a normal distribution. They selected the twenty most relevant chosen predictors by using Gradient Boosting Machine. The number of "twenty" variables was chosen after a sensitivity analysis by comparing the models with R^2 and AIC. As a result of the variables with the most influence were related to health, restaurants, bookstores, banks, bus stations and car dealers, among others. Positive impacts had the densities of banks, car dealers, health, restaurants (except from 12-16h), while negative impacts presented the bus stations and the books stores (except 4-8h).

As another example, Schmöller and Bogenberger (2014) considered the trips of FFCS and a hybrid solution (HCS) where shared cars can be parked only in specific areas in Munich. Their visual analysis showed that areas with more trips are related to the old city border, distance to the city center and the two most prominent universities in Munich. According to the socio-demographic, FFCS has a higher usage in zones with a higher ratio of young inhabitants. However, the weather was not correlated to the bookings. Willing et al. (2017) developed a model correlating the points of interest (POIs) (e.g., hospitals, banks, restaurants.) with the demand of FFCS in the city of Amsterdam. From 94 POIs they selected the 20 most important for different time intervals using a gradient boosting machine (GBM) method. Finally, they applied a generalized linear model (GLM) to predict the demand.

Seign et al. (2015) developed a model to forecast FFCS hot-spots after success factors. The study area was divided into tracts or segments according to districts ("Wohnquartier") including the bookings per km². The model was build after a Pearson's correlation including population density, restaurants and hotels density, distance to the city center, and rent prices. They concluded indicating that research is required to reveal possible gaps in public transport that could be filled by shared vehicles.

2.3.2 Factors affecting the deployment of bike sharing systems

Table 2.4 summarizes some studies that researched about the factor influencing station based bike sharing systems. It includes the cities from all over the world with different sizes. The most common regression method is ordinary least squares. The dependent variable mostly used is the logarithm of the number or rates of arrivals and departures. To assess the models the indexes mostly used are log-likelihood (LL), R^2 and AIC-BIC. Furthermore, Table 2.5 shows the resulting most important exogenous and endogenous factors in SBBS. The following section explains in detail some of these models.

Additionally, Table 2.6 indicates the buffer distances that some studies adapted as zones of influence of the stations. This buffer helps to select the spatial variables affecting a station. The criteria of the authors is that these values are appropriate walking distance to and from stations to rent a bike or to develop an activity.

Chardon et al. (2017) studied the trips per day per bicycle (TDB) in ca city level in 75 SBBS systems in Europe, Israel, United States, Canada, Brazil and Australia. They stated that the average TBD is in a range from 0.22 to 8.4. They used a robust regression to build the model because it reduces the impact of outliers and the log of the TBD as dependent variables. The independent variables were the operator's attributes, the compactness, the weather, the transportation infrastructure and the geography. They used a buffer of 300 meters to measure the coverage area. They selected the best model after mixing different model structures by AIC, BIC, R^2 and log likelihood. Helmet requirement, low temperature, high wind speed, low number of docks at stations decrease the performance of the systems, while high population, high cycling infrastructure, and stations density increased it.

Table 2.4: Related work concerning factors affecting SBBS

Author	City Name	Spatial Scale	Dependent Variables	Model Type	Model Assessment
Chardon et al. (2017)	75 cities world wide	City	Log trip/day per bicycle	Robust linear regression	AIC, BIC, R^2 , LL
Zhao et al. (2014)	69 cities in China	City	Log Daily use + turnover rate	OLS and partial least squares	R^2
Faghih-Imani and Eluru (2016)	New York	Station	Log arrivals and departures rates	Pooled linear regression + spatial and temporal lagged	AIC, BIC, # parameters, LL
Noland et al. (2016)	New York	Station	Number of trips	Negative binomial regression	AIC, LL, R^2
Wang et al. (2015)	Minneapolis + St. Paul	Station	Log number of trips	OLS	R^2
El-Assi et al. (2017)	Toronto	Station	Log number of trips	OLS	R^2
Faghih-Imani et al. (2014)	Montreal	Station	Hourly arrivals and departures	OLS	LL, BIC
Tran et al. (2015)	Lyon	Station	Arrivals or departures per hour	Robust linear regression	R^2
Faghih-Imani et al. (2017)	Barcelona + Seville	Subcity district	Log arrivals and departures rates	OLS + Auto regressive moving average	LL
Mattson and Godavarthy (2017)	Fargo	Station	Log (rides per day+1)	OLS	R^2
Caulfield et al. (2017)	Cork	Station	Probability trip travel time falls in a category	Logistic regression	R^2

Zhao et al. (2014) correlated the logarithm of ridership and turn over rate using data of 69 SSBS systems in China with urban features and system characteristics. The regressions were ordinary least squares (OLS) and partial least squares (PLS). They omitted correlated variables by limiting the variance inflation factor (VIF) to 10 and the Pearson correlation with a threshold of 0.7. They ranked the variables using the variable importance for projection (VIP) index. Results indicated that ridership and turn over rate increased with the population, government expenditure, users of the system and docking stations, personal credit cards and universal cards. The number of bicycles increased the ridership but lowered the turn over rate.

Tran et al. (2015) did a robust linear regression model of station-based bike sharing (SBBS) aggregated hourly arrivals and departures in Lyon. They did not consider trips less than 3 minutes and longer than three hours and during July and August. The data were calculated in a buffer of 300m after a sensitivity analysis of 200m, 300m, 400m. The external variables used were related to public transport, socioeconomic, topography, bike sharing network and leisure activities. The results show that long-term members used the system for commuting trips, but short-term members used it for leisure activities. The principal factors affecting positively the flows were the rain stations, restaurant, cinema and embankment roads, while the altitude presented a negative influence. Furthermore, the population density showed a positive effect in

Table 2.5: Factors affecting the deployment of SBBS

Category	Variable	A	B	C	D	E	F	G	H	I	J
Scheme Design	Stations density	✓						✓	✓		
	Docks per station	✓	✓		✓		✓	✓	✓	✓	✓
City size	City Population	✓	✓								
Demography	Population density			✓				✓	✓	✓	✓
	Jobs density			✓		✓		✓	✓		
Topography	Altitude								✓	✓	
Existing Infrastructure	Cycling infrastructure	✓		✓							
	Railways length			✓							
	Subway stations			✓	✓		✓		✓	✓	
	Rail stations								✓		
	Universities						✓	✓			✓
	Student residence								✓		
	Restaurants			✓		✓		✓	✓		
	Cinema								✓		
	Distance to CBD						✓		✓		
	Number of business						✓		✓		✓
	Parks			✓							
	Residential land use					✓					
	Parking land use					✓					
Distance to water body						✓					

A= Chardon et al. (2017), B= Zhao et al. (2014), C= Faghih-Imani and Eluru (2016), D= Noland et al. (2016), E= Wang et al. (2015), F= El-Assi et al. (2017), G= Faghih-Imani et al. (2014), H= Tran et al. (2015), I= Faghih-Imani et al. (2017), J= Mattson and Godavarthy (2017)

Table 2.6: Distance of influence from the stations

Author	Distance of influence [m]
Schmöller and Bogenberger (2014)	400
Noland et al. (2016)	400
Wang et al. (2015)	400
Tran et al. (2015)	300
Chardon et al. (2017)	300
Faghih-Imani et al. (2014)	250
El-Assi et al. (2017)	200

the morning and the number of jobs with a positive impact in the afternoon.

On the other hand, Faghih-Imani and Eluru (2016) correlated the hourly arrivals and departures for one month in a station based bike sharing system "CitiBike" in New York with temporal, spatial and weather variables. Spatial variables included population density at a zip-code level, in a 250 meters radius buffer they considered the transportation system attributes, restaurants, and area of parks. They implemented 3 groups of models. One of them was a pooled linear regression considering spatial panel models, the dependent variables where the log-normalized arrivals and departures rates. Then, for the other two, they added spatially

lagged dependent variable (spatial lag model) and spatial autocorrelation process in the error term (spatial error model). They compared the different models by the number of parameters, AIC, BIC and the log-likelihood at convergence. To validate the model they calculated Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) from the data of the week after the training data. Models showed a better performance at PM and evening. The best-fitted model was the spatial lagged model with both temporally and spatially lagged variables. They concluded that the fit of the model improved significantly by adding temporally and spatially lagged dependent variables. The temperature did not have a significant effect, but the rain did. The length of bicycle routes, the presence of subway stations, the area of parks on weekends, number of restaurants increased the usage of the system, while the length of railways decreased it. Population and jobs density showed the daily worked-based trips. The arrivals and departures of one station were correlated to those of the neighboring stations. They concluded that the land use variables should also be included in further researches.

[Caulfield et al. \(2017\)](#) analyzed the usage patterns of SBBS in Cork, Ireland. The probability that a trip falls into a time interval was examined by a logistic regression model. This study took into trip variables (travel time, time of the day and the weather) and bike station's characteristics (the frequency of usage,). The principal results are that the shortest journeys are realized by the most frequent users, at the busiest stations, at the AM peak, and through the most popular routes. The shortest trips are carried out during the weekdays, in high precipitation conditions, and warmer and brighter days. Finally, they suggested more research on the understanding of the difference between these variables in small and large cities.

In respect to exogenous factors influencing free floating bike sharing, [Reiss and Bogenberger \(2017\)](#) showed spatial and temporal mobility patterns in a free-floating bike sharing (FFBS) system in Munich. Precipitation events led to fewer bookings. But major events, like concerts or sports games, took to an imbalance of the network due to a high impact on the demand.

Part II

Methodological framework

Methodology overview

In this part of the thesis, the proposed and applied methodological framework is presented. In order to reach the objectives of the research, the a automated methodology is divided into three main components that are required to adequately model the demand for shared mobility: 1) Automated data collection, 2) Automated data analysis and processing 3) Automated Model building and selection (Figure 2.5). It is worth to say that every step in the methodology belongs to one script and it is automatically performed

The proposed data collection methodology is initiated by the collection of the variables from one or several car or bike sharing systems. After an exploratory data analysis, time intervals are set as the time unit, and influence zones are delineated as the spatial unit. The daily averages of the arrivals and departures are aggregated by time intervals and spatial influence zones. Second, the independent variables that potentially describe the system are extracted from open- source datasets which are intersected with the zones of influence (defined in the pertinent section). The two dataset are combined based on their common spatial reference (zone of influence) in order to estimate models that will allow to predict the demand.

The independent spatial variables are combined with the zones of influence to proceed with the calculation of two indicators per variable: one representing the dispersion (distance) and the other measuring the quantity (density or presence). The variables that are not relevant for the study are removed, as are the variables that are not included in the whole studied systems. We want an outcome with variables of a general city without any specific factors.

Before the models are built, a collinearity analysis between the data has to be carried out to remove redundant variables. Then, different types of models are created to select the one that better fits the dataset. Since the models might present problems, such as outliers and heteroscedasticity, among others, some transformations of the data set might be required. The variables that influence the models' outcome the most are aggregated, and the most representative are selected. Finally, the best model is chosen after indirect comparison methods, a validation test and a k-fold cross-validation.

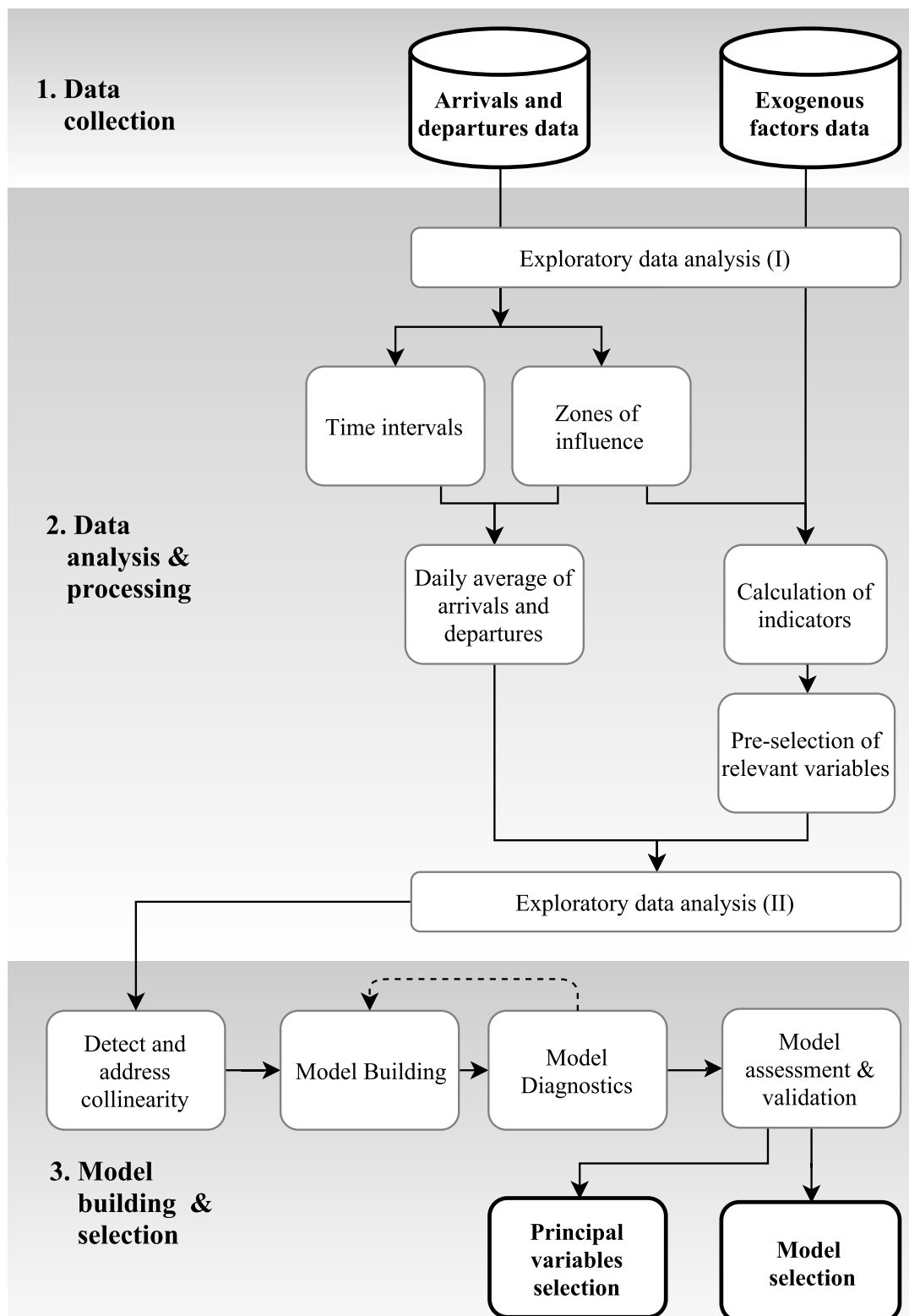


Figure 2.5: Methodological framework

Chapter 3

Data collection, analysis, and processing

This section includes the method to collect, analyze and process the dataset of the proposed methodology. First, the data collection and the requirements of the dependent and independent variables are presented. Then, an exploratory data analysis is needed after collecting the data to define time intervals, zones of influence and to calculate indicators of the spatial independent variables. Finally, the variables that are going to be part of the construction of the models are pre-selected and analyzed.

3.1 Data collection

In order to start the automated methodology the open-source data has to be downloaded. The data needed to reach the objectives of the research have to be collected from one or more than one bike sharing or car sharing systems. Given the fact that the demand has to be modeled, the dependent variables are the average daily arrivals and departures of the rentals of these systems, and the independent variables are the exogenous factors. Also, given the fact that the common aspect of the dependent and independent variables is the zone of influence it is imperative that all data have a spatial reference: coordinates (e.g., latitude and longitude). Finally, each arrival or departure has to include the origin and destination time respectively: the date (day/month/year) and the hour (HH:MM:SS).

The evolution of ICT as well as the open-source data movement has changed the landscape of data availability in almost every sector and has allowed for the emergence of data sharing platforms. These platforms are in many cases collective efforts that provide data of different typology and content. The independent variables are the exogenous factors listed in Section 2.3. Apparently, given the vast quantities of data, a pre-selection is required, that is mainly bounded by the required characteristics of the proposed methodology.

Therefore, variables that are not relevant for the study have to be removed according to the criteria and expertise of the author. It is worth to say that this is the only manual step in the script that is required. Four criteria are considered to remove the variables:

- a) Relevance on the study objectives (e.g. remove vending machines)
- b) Accuracy of the data (e.g. unclassified streets).
- c) Repeated data (e.g. points of interest as spatial points and points of interest as spatial polygons).
- d) Unnecessary subdivisions for the purpose of the study (e.g. unnecessary subdivisions of track roads).

Furthermore, the remaining variables might have an unneeded large spatial scale (e.g., country scale). To improve the speed of the data processing, information outside the relevant spatial area has to be removed. Therefore, the variables are clipped within a two-kilometer buffer from the perimeter of the rentals' location. This distance can be variable, but we consider this a conservative size for the study purposes.

3.2 Time intervals

Time intervals are time periods which are going to function as time unit for the dependent variables. The pre-requisite of the rentals to have the date and hour of arrivals and departures guide the assignment of arrivals and departures in five different temporal units:

- Hour (e.g., 12)
- Day of the week (e.g., Monday)
- Type of day (weekday or workday)
- Month (e.g., June)
- Year (e.g., 2014)

As the arrivals and departures are aggregated by the same time unit, an exploratory data analysis (EDA) can be performed. For example, scatterplots, bar plots, and boxplots should be used to analyze different temporal patterns and compare the rentals' distribution in the different cities.

Six day-intervals are set to each arrival and departure. These intervals are defined according to the hourly distribution of arrivals and departures and the time of the day (morning, afternoon and night). They are built by subdividing the time of the day into peak and off-peak:

- Morning peak
- Morning off-peak
- Afternoon peak
- Afternoon off-peak
- Night peak
- Night off-peak

Then, the time units are aggregated as *day of week + time interval* (Saturday Afternoon off-peak), and *type of day + time interval* (e.g., Workday Morning peak). This means two extra time units are assigned to each arrival and departure.

3.3 Zones of influence

The cities or systems included in the study have to be divided into zones of influence, which essentially represent the same spatial scale for dependent and independent variables. This allows the establishment of a relationship between the input and output, thus the derivation of regression models. Dependent and independent variables are aggregated and assigned to a corresponding zone of influence. Station-based and free-floating-based use two different methods to delineate the zone boundaries.

– **Station-based systems** – For station-based systems, the boundaries of the zones correspond to the limits of the area of influence of each station, as well as, natural and human-made barriers. The boundaries of these areas can be limited by:

- a) *Voronoi diagrams* (also called Thiessen polygons or Dirichlet tessellations). They correspond to a subdivision of an area into Voronoi cells. The limits of these cells are all the points that have the same distance to the two nearest stations (Voronoi, 1908).
- b) *Station's buffer*. This area represents the furthest that a user is typically willing to walk to access a shared vehicle or to perform an activity after returning the vehicle. Buffers with a radius of influence from 200 to 400 meters from the stations are commonly used in the literature (see Table 2.6). However, a sensitivity analysis is recommended to find the distance that better captures the phenomenon.
- c) *Postal code zones*. They are considered barriers since their delineation is usually by high hierarchy roads. Usually, these roads reduce the access to a station.
- d) *Riverbanks*. They are considered as limits for the zones since they are barriers for walking or cycling to the stations.

The procedure to delineate the zones of influence (Figure 3.1) starts by creating a bounding box surrounding the stations. This box includes a conservative distance of a buffer (e.g., one kilometer) and it will be the boundary of the Voronoi diagrams. To calculate the Voronoi polygons or tessellations for the stations' locations, we use the function `voronoi.polygon` included in the package `SDraw` in the statistical programming language R. Finally, the zones of influence are the product of intersection the Voronoi polygons, the influence buffer, the postal code zones and the riverbanks.

– **Free-floating-based systems** – Since free-floating-based systems do not have stations, the arrivals and departures can be grouped into clusters. The centroid of each cluster would represent a station. Then, to configure the boundaries of the zones of influence, the same procedure for the station-based systems can be followed. Some clusters techniques that can be used are:

1. *Partitioning methods* (e.g., k-means). K-means (MacQueen, 1967) initializes by locating k centers and assigning each data point to a center. Then, the center's locations are recalculated by the mean of each cluster group. This process is repeated until the centers converge. This algorithm can be implemented by the function `kmeans` in R. The main input arguments required for this function is the number of centers k .
2. *Grid-based methods*. Grid clustering is based on dividing the area of study into a grid system with $m * n$ -sized cells.
3. *Density-based methods* (e.g., density-based spatial clustering of applications with noise (DBSCAN)). DBSCAN is an algorithm that clusters spatial data in arbitrary shapes (Ester et al., 1996). It based on finding the closest neighborhoods of each point limited by a minimum number of points. This method can be implemented by using the function `dbscan` in the programming language R. It requires the minimum number of points and the maximum distance the between neighbors as input arguments.

Finally, a sensitivity analysis helps to determine the parameters that would best perform, and also, to estimate the best clustering method that fits the data of the study.

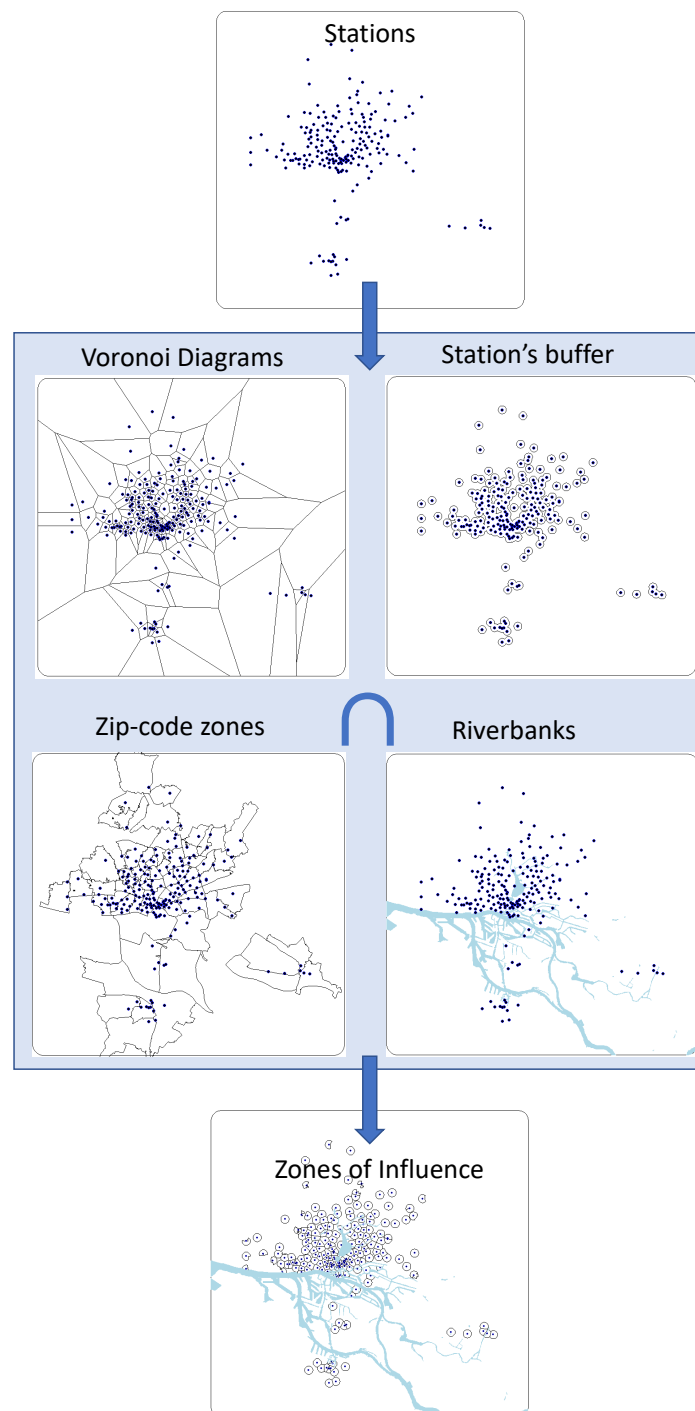


Figure 3.1: Delineation of the zones of influence in station-based shared systems

3.4 Dependent variable: average daily arrivals and departures

Average daily arrivals and departures are the dependent variables for the models. In this study, we also call them observations or responses. Arrivals and departures are assigned and the aggregated to the corresponding zone of influence and to the different time units. To calculate the daily average, arrivals and departures are divided by the total number of days according to the time unit (e.g., total workdays, total Sundays).

Since the time units might be correlated with each other (e.g., Tuesday morning and Wednesday morning), not all of them are needed to be modeled. Therefore, a Pearson correlation test is carried out to determine the time units that differ the most. Only the time units with the arrivals and departures that are less correlated are taken into account.

3.5 Calculation of the indicators for the spatial variables

Each independent variable has to be assigned to a zone of influence. If a variable is unique for each city of the study (e.g., population, time of operation, total number of stations), all the zones of influence within a city have the same attribute. However, if the variable differs from each zone of influence, we need to set a dimension or indicator to measure the weight of each variable.

Two indicators are assigned for each spatial variable. One indicator represents the distance from a station to the variable and other the quantity or presence in a zone of influence. Each variable presents a different spatial impact within the city. Therefore, the indicators differ for each variable. For example, some infrastructure influences the arrivals and departures of the whole city as the location of universities or the city center. However, other types of variables affect the destination just within the zone of influence and not at the city level, for instance, a bus station. Another example is the irrelevancy of the presence of a single tree in a zone of influence. However, the trees' density might be relevant in the study because it might mean the presence of a green area.

After these considerations, four possible families of indicators are derived per zone of influence:

- *Density*: Frequency per area unit.
- *InArea*: Boolean 1 if the variable is present in the zone of influence, or 0 if it is not.
- *Distance_min*: Minimum distance from a station to a spatial variable.
- *Distance_min_all*: Minimum distance from a station to all the variables within a city.

The above described process might yield to a vast number of variables (depending on the source) that are considered candidates for the extraction of a model that can describe the relationship between the input and output data. However, given issues commonly met in the inclusion of a large number of variables in the model development phase (presented in Section) it is important to determine which indicator fits better each variable.

Therefore, we assume that if the variables are relatively few in a city (present in less than the 3% of all zones of influence, e.g., universities) then, they are assigned the indicator *Distance_min_all*, any other way the indicator would be *Distance_min*. Moreover, if a variable behaves with a very different distribution or spread (standard deviation) in the city less than threshold, the indicator *InArea* is taken into account. Otherwise, the indicator *Density* is considered. For example, if the most of the zones of influence have the same variables amount of variables, we care about the presence of those variables but if they are very different distributed in the city we care about their density.

Once the exogenous factors are intersected with the zones of influence, the procedure continues by calculating the distance from the stations (centroids) to all the features. Then, *Distance_min* is taken as the minimum distance from a station (centroid) to a variable inside the zone of influence. If a variable is not inside the zone, the indicator *Distance_min* is assigned to a significant large distance of 99999 to lower its cost (since zero is the highest cost). If more than a low percentage (e.g., 3%) of the zones of influence have a value of *Distance_min* = 99999 then the indicator *Distance_min_all* is taken into account.

Furthermore, the spatial variables have to be classified into points, lines or polygons. Lines and polygons after their two dimensions, only the indicator *Density* is considered. Nevertheless, for points, the standard deviation of their frequency in the zones of influence is used to determine if the features are equally spread in the city or not. If their rates are uniform, this means that just the presence matters and not the quantity. The issue is to determine which standard deviation is a suitable threshold. Therefore, a sensitivity analysis has to be carried out to determine the best value of the threshold of the standard deviation. Finally, the more equally distributed variables are assigned the indicator *InArea*, otherwise *Density*. Figure 3.2 shows a summary of the process for calculating the indicators.

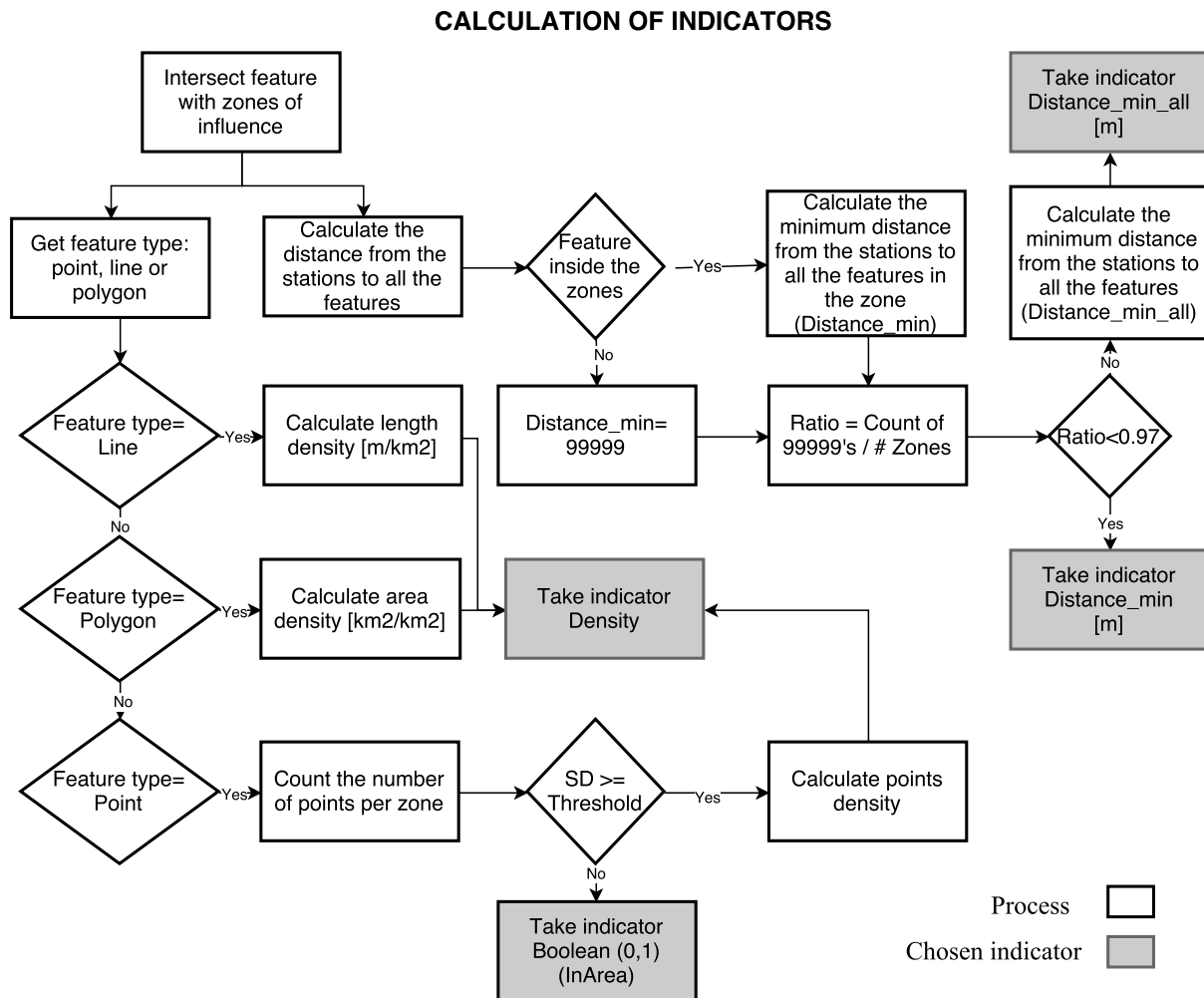


Figure 3.2: Calculation of indicators

3.6 Pre-selection of variables

In this stage, the independent variables that coincide between the cities of study are chosen. The outcome models will not consider individual cities, but they are going to treat all zones as one system. The model's objective is to predict the arrivals and departures in a city or a system without mattering its size. Therefore, the model will not take into account specific factors of one city, i.e., the considered variables have to be the same in the whole cities of the study and to present similar behavior (indicators). Thus, we have high probabilities that the model to predict the offer or demand includes the variables in the city of study. For example, the

distance to the beach in a not coastal city in the test sample would be read as 99999, affecting the model. But in reality, the absence of a beach does not affect the arrivals or departures.

3.7 Exploratory data analysis (EDA)

Finally, after the variables that are going to take part of the study are pre-selected, an exploratory data analysis (EDA) is realized. As shown in Table 2.4, in the literature is very common to use the logarithm of rates of arrivals and departures as dependent variables. After an EDA, we can check this assumption that our dependent variable is going to behave better in our models by considering its logarithmic transformation. Additionally, we can detect some mistakes in the dataset as outliers and further relationships among the independent and dependent variables.

For an EDA, we can use the following techniques:

- Central tendency values: mean, median.
- Spread: standard deviation.
- Maximum and minimum.
- Percentiles: 25
- Histograms
- Boxplots
- Scatterplots
- Correlation matrices: Pearson, Spearman

The first four techniques can be deployed by the *summarize* function in the programming language R. Histograms, boxplots and scatterplots can be printed using the function *ggpairs* of the package *GGally*. Finally, the function *cor* from the package *stats* helps us to generate correlation matrices using the attribute *"method"* equal to *"pearson"* or *"spearman"*.

Chapter 4

Model building and selection

This section details the process to build the models that correlate the average daily arrivals and departures with the exogenous factors and how to select the model that best fits the dataset. Furthermore, to reach the objectives of the research, the primary variables affecting the deployment of car and bike sharing are selected (see Figure 4.1). First, we detect and address collinearity. Then we build of three different type of regression models for each time unit: stepwise regression, generalized linear models, and gradient boosting machine. After solving the problems detected, the primary variables and the better model are selected. The best is the one that fits better the data after comparing them with indirect methods, model's validation and overfit control.

4.1 Detecting and addressing collinearity

Collinearity has to be tested before building a model in order to not have lower standard errors and and more stability specially in OLS (Mack, 2016) . To detect collinearity, the three criteria mentioned in Section 2.2.5 are taken into account. Then, if collinearity is present, two techniques can be used to remove high correlated variables :

1. **Direct elimination.** This technique is based on removing from the dataset the most correlated variables after a Pearson correlation analysis. A value of 0.7 is taken as threshold since it is a reasonable limit as mentioned in Zhao et al. (2014). First, we have to build a correlation matrix with the function "cor" from the programming language R. If the correlation value between two variables is higher than 0.7 then, the variable with the greater total sum of the correlation coefficients from the whole dataset remains and the lower is removed.
2. **VIF criterion-based.** This technique is based on the VIF criterion to detect collinearity (see Section 2.2.5). It starts calculating the VIF for each variable and the variable with the highest VIF is removed. The VIF of each variable is calculated again for the remaining variables and the one with the greatest VIF is again removed. This process is repeated until no variable has a VIF higher than 5.

To select the best technique to address collinearity, a sensitivity analysis has to be carried out. A ordinary least squares (OLS) method can be used on the remaining variables to decide which technique works better on the dataset of the thesis.

4.2 Model building

To estimate or predict arrivals and departures or to select the most relevant exogenous variables in a shared transportation system, modeling procedures are applied. The goal is to build a

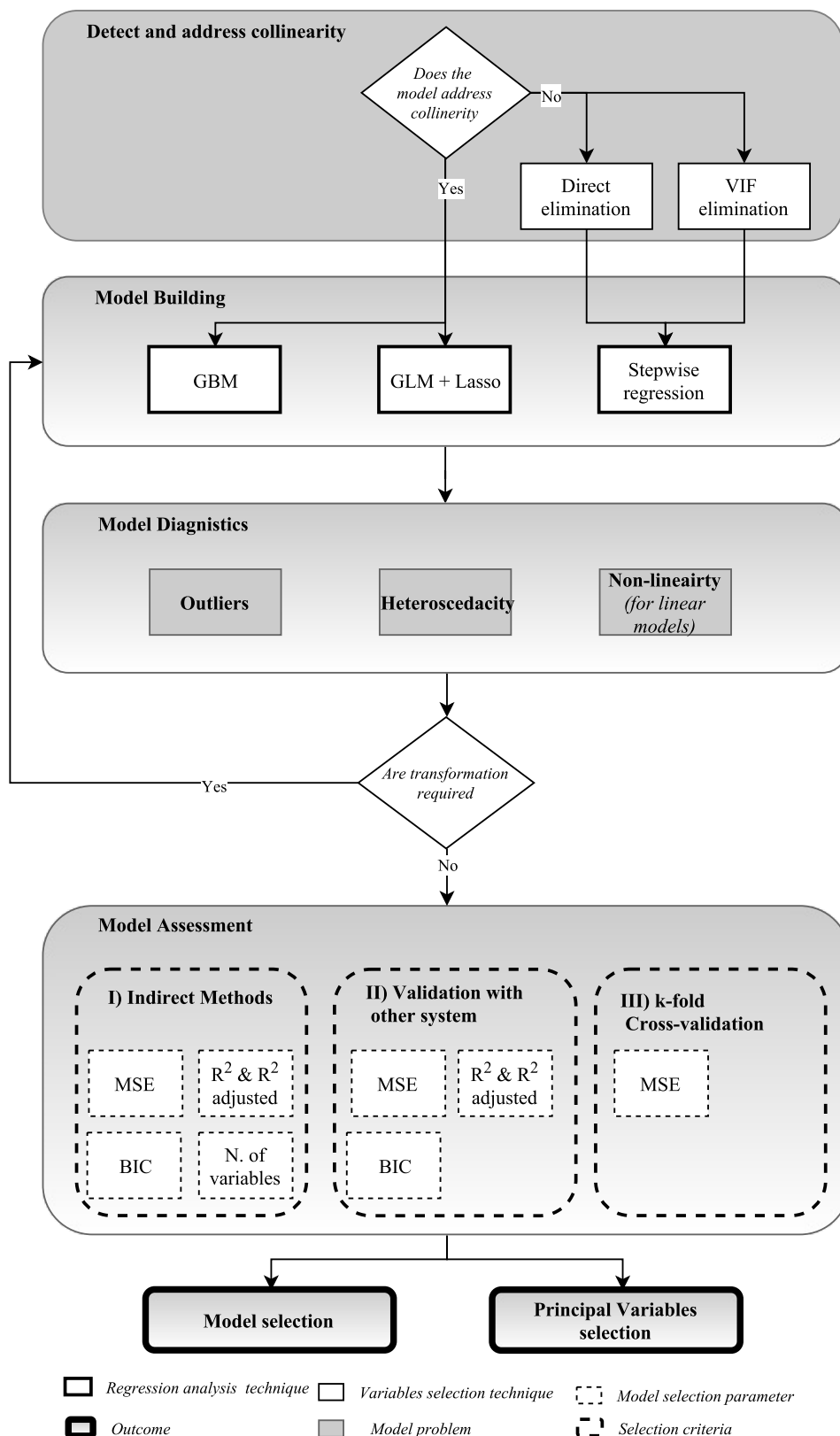


Figure 4.1: Model building and selection

regression model that better fits the data set and also do a classification of the most influential variables. Therefore, the regression methods to build the models have to be chosen. Two criteria are considered for this selection: 1) the models have to deal with either linear or nonlinear models

and 2) a variable selection or variable ranking has to be implied in the method.

The dataset is assumed to either behave linear or nonlinear. So, the regression methods selected to identify relationships between dependent and independent variables have to analyze both possibilities: linearity and nonlinearity. Moreover, the methods are required to perform simple models with the minimum number of variables as possible without losing their fitting performance. An additional requirement of the methods is that they include a variable selection or categorize the variables for computational efficiency, avoiding errors, and simplicity. Computational efficiency means that if in the models a variable selection is included, usually functions from programming languages have the best script with less computational effort. Avoiding errors is needed that if further programming has to be done some assumptions might be done or coding mistakes can be realized that might reduce the performance of the models. And last but not least, simplicity means that the final script for the methodology is simpler to code and to understand.

Several regression methods could have been chosen. However, in this methodology, just three methods are presented mainly because of the scope of the thesis. The three types of regression models that were chosen to correlate the dependent and independent variables were: two linear models (stepwise regression and GLM) and one decision tree regression model (GBM). OLS with a stepwise regression for variables selection was chosen because OLS is highly implemented in the literature (see Table 2.5) and secondly because with the stepwise regression technique, a variable selection is implied in the model. GLM is commonly used when OLS is not appropriate for linear models, and furthermore, through the lasso technique, it performs a variable selection. Finally, decision trees regression are also considered because it widely used and even in the case that the dataset might perform better with a nonlinear model. The gradient boosting approach is taken into account since it helps decision trees to improve their prediction. It does not perform a variable selection, but it does a variable ranking. An additional reason because GLM and GBM were selected, was because their good performance on a similar approach used by [Willing et al. \(2017\)](#).

On the other hand, PCA is an example of an excluded model. In the literature review, the PCA technique was mentioned as a method to address collinearity and regression. However, it is not used since it does not match the criteria of doing a direct variable selection or a ranking of them, it just addresses the problem of collinearity. Even though, using some techniques as choosing the main principal components and their included variables can be used.

– **Stepwise regression** – This regression model does not deal with collinearity. After the detecting and addressing collinearity, a stepwise regression can be performed. The method of both directions is chosen because can eliminate irrelevant variables in further steps. The function *stepAIC* included the *MASS* package from the R statistical programming language helps to develop these models. This function requires principally a formula (Equation 2.9), a direction (backward,forward,both) and a k parameter. The direction is set as "both" because if a nonsignificant variable enters early in the process, it might be eliminated later. k is a parameters that indicates if we want to use AIC ($k = 2$) or BIC ($k = \log(\#observations)$). BIC is selected as a value to compare each step in the model because AIC usually presents higher instability and it produces more complex models (see Section 2.2.6).

– **Generalized linear model + lasso** – For generalized linear models, we chose lasso as shrinkage method because it has the possibility to shrink coefficients to zero. Therefore, it produces a more parsimonious model with fewer variables and less chances of overfitting. The function *glmnet* from the package *glmnet* included in R statistical programming language is used to build this type of models. The principal input arguments of the function are the dependent and independent variables, the shrinkage coefficient λ , the distribution function of the error and a parameter α . If α is equal to one than the GLM uses "lasso" as shrinkage

method.

Furthermore, two sensitivity analysis have to be performed : 1) to determine the best distribution of the error ("gaussian", "binomial", "poisson", "multinomial", "cox", "mgaussian"), 2) to choose the best shrinkage factor λ . For the first analysis, the model has to be run for all the admissible distributions to see which one fits better the data. The results of the different models can be compared with a indirect method as R^2 adjusted.

To choose the best λ is more complicated because a k-fold cross-validation test has to be performed. The function `cv.glmnet` from the package `glmnet` does the validation with the same input arguments of the previous function plus the attribute `nfolds` (number of folds for the cross validation). There are two principal results from the function `cv.glmnet`: 1) λ_{min} and 2) λ_{1se} . λ_{min} is the λ that gives the lowest MSE, however, λ_{1se} is "largest value of λ such that error is within one standard error of the minimum". λ_{1se} is chosen because it selects fewer variables (simpler model) and it reduces the risk of overfitting. Another result of `cv.glmnet` is a plot representing the $\log(\lambda)$ vs. the MSE (see Figure 4.2). The two vertical lines represent λ_{min} and λ_{1se} and the values above the plot are the number of selected variables.

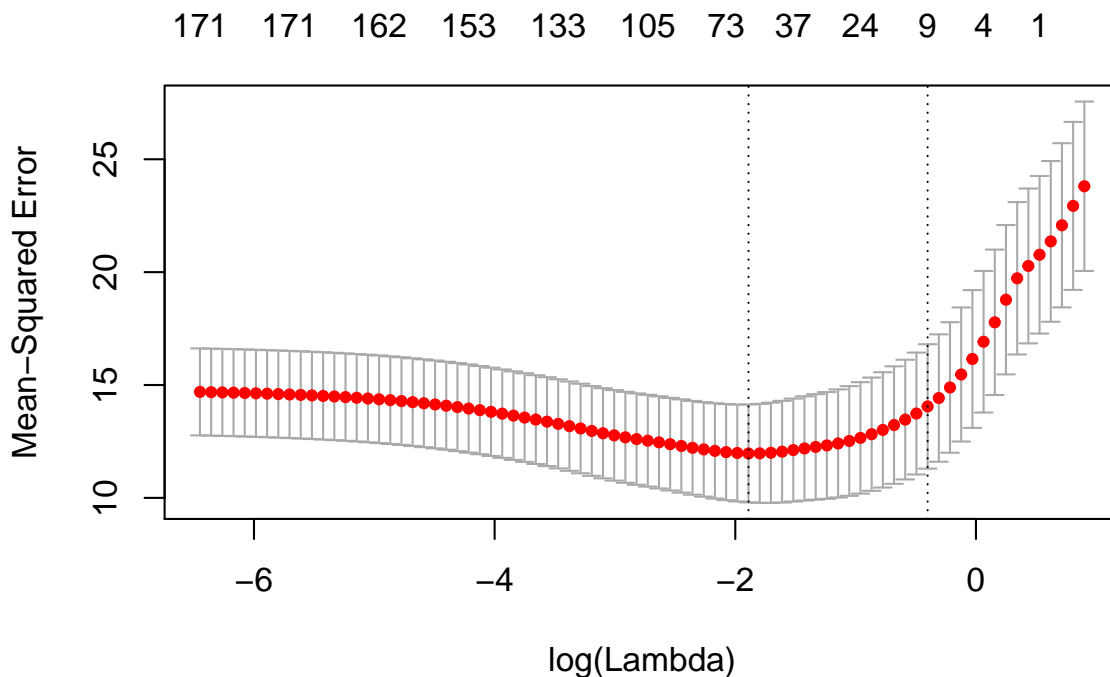


Figure 4.2: Example k-folds cross-validation test to find the best value of λ

– **Gradient Boosting Machine** – The function `gbm` from the package `gbm` is used to develop GBM models. It uses the input arguments described in Section 2.2.8: `distribution` (loss function), `n.tree` (number of iterations), `interaction.depth` (terminal nodes of each tree) and shrinkage (λ_{gbm}). A sensitivity analysis has to be carried out to determine the loss function that better fits the dataset. Six terminal nodes are selected of each tree as suggested in the literature (Section 2.2.8). Even though, a very small λ_{gbm} requires a high computational effort, a very conservative $\lambda_{gbm} = 0.001$ is adopted to have better results. To chose the better number of trees, a k-fold cross-validation is generated with 5 folds. This number is the minimum value recommended in the literature, and it is used to reduce the computational effort. The cross-validation is carry out by introducing the attribute `cv.folds = 5` in the function `gbm`. Then,

the function `gbm.perf` displays the best the number of trees after setting as input arguments the the `method` as `cv` and the set of models resulted from the cross-validation (see Figure 4.3). In addition to the resulting model, `gbm` provides a ranking list of the variables with their relative influence. The relative influence is normalized to sum one hundred. For the "gaussian" distribution the relative influence is the percentage that each variable contributes to reduce the squared error. For the other loss function, it describes the relative influence of each variables reducing the loss function.

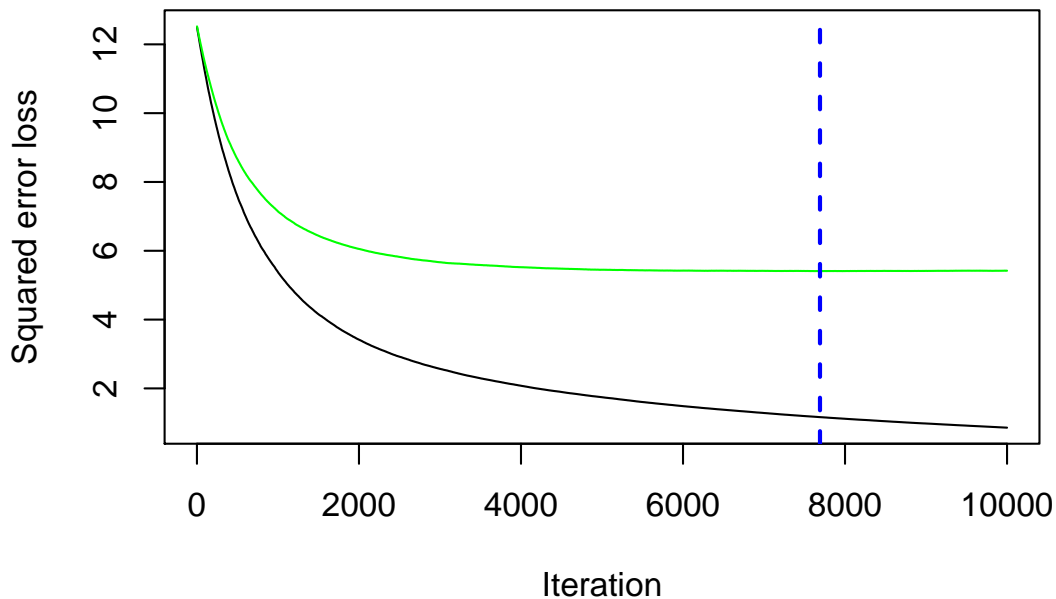


Figure 4.3: Example k-folds cross-validation test to find the best number of iterations (trees)

4.3 Model diagnostics

Stepwise regression and GLM have to assume a linear response-predictor relationship and homoscedasticity. Therefore, we plot the predicted response vs. the MSE to analyze if they have problems of heteroscedasticity and nonlinear response-predictor relationship. We can follow the same procedure for GBM to examine heteroscedasticity. The function `predict` from the package `stats` can be implemented to calculate the predicted response. For GBM and GLM, this function requires two input arguments: 1) the resulting model and 2) the parameters used to build this model. For the stepwise regression, only the resulting model is required.

If the mentioned problems are detected, transformations of the variables have to be considered as discussed in Section 2.2.4. Most common transformations are: power transforms the square root, logarithmic, inverse, exponential, arcsine (Bishara and Hittner, 2012). One possible solution for heteroscedasticity is BoxCox transformation (see Section 2.2.4). The function `boxcox` from the package `MASS` in the programming language R is required to find the value λ that maximizes the SSR (see Figure 4.4). The main attribute of this function is a linear model built with the function `lm` by using the variables that are going to be part of the model. Once the boxcox or other transformations are performed, the models have to be rebuilt. Finally, to

identify outliers, the function *outlierTest* from the package *car* select them. Once they are removed, the model has to be built again.

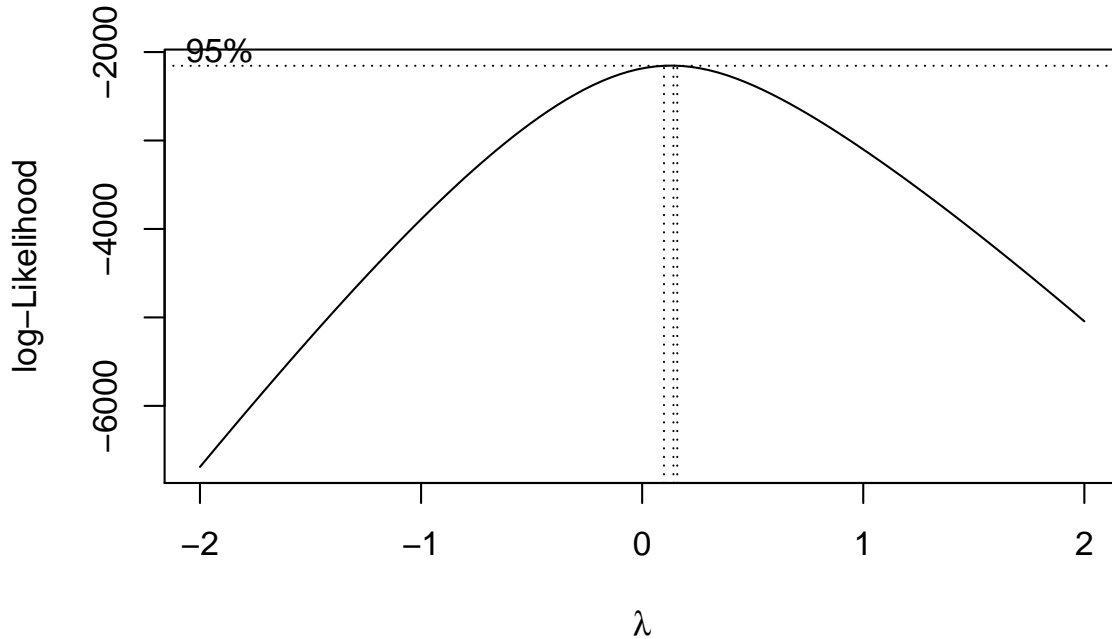


Figure 4.4: Example of finding the λ value for the Boxcox transformation

4.4 Model assessment and selection

This stage is the final of the thesis' methodology. The aim of this section is to analyze which models fit better the data and include the lowest number of variables. Therefore, all the built models are compared after an assessment of them. Three types of criteria are used to assess the built models:

Indirect methods: lowest number of predictors, lowest *MSE*, lowest *BIC*, and greatest R^2 and R_{adj}^2 .

Best validation results We want to select the model that adequately predicts the demand on a validation dataset (e.g., for another city or system). We have to do the same process with the cities within the study to calculate the dependent and independent variables for the validation. We predict the dependent variable, and then the parameters R^2 and *MSE* are considered to measure the performance of the validation. *BIC* and R_{adj}^2 can also take part of the assessment, but we need a city or system with a higher number of stations (centroids) than the total observations used to build the models. To carry out the validation, the function *predict* from R can be used to estimate the predicted dependent variable to validate the different models.

Lowest overfit To check if the model is overfitting problems, we can do a k-fold cross-validation. If the *MSE* calculated for each fold changes drastically, we assume that it might be an overfitting issue. Because of computational effort, we do not recommend more than five

fold. In conclusion, the model with relative less MSE's differences is the one that overfits the less.

However, the three different criteria do not have the same importance. The validation has the highest weight because it is a direct way to measure the performance of the models. The second is the overfitting test because even if you have better values on the indirect methods if there is a problem of overfitting the model will not have the best performance. Finally, when these three criteria are analyzed, the model that best fitted the criteria is selected for each time unit.

4.5 Principal variables selection

We assume that the variables that influence the most the deployment of a car or bike sharing systems are the variables selected from stepwise regression and GLM or the first variables ranked from the GBM method. For stepwise regression and GLM, the selection of variables is not an issue because it is already implicit in the models' output. However, GBM does not perform a variable selection but ranking list according to the influence of the variables. Therefore, to select the most important variables, a validation test has to be performed. We build a GBM model with only with the variable with the highest influence and we calculate the resulting MSE. Then, the process is repeated by adding the second variable with more influence. It continues until the last variable is included in the model. The number of variables are selected where there is not a significant reduce of the MSE. A plot of the number of variables vs the MSE might help to identify the threshold.

As the most significant variables are selected for each model and each time unit, we count how many times a variable is cataloged as selected. To conclude, the variables with higher frequency are the ones that we classify as "variables with higher influence" on the respective systems.

Part III
Case of Study

Chapter 5

Area of implementation

The present chapter summarizes the area of implementation of the automated methodology describe in the previous part. The case study for this research are six SBBS systems of the operator Call a Bike in six cities in Germany: Hamburg, Frankfurt, Kassel, Stuttgart, Darmstadt, and Marburg. This chapter summarizes the main characteristics and shared mobility in Germany and also the principal aspects of these six cities. Finally, we explain how Call a Bike works in general and in the studied cities.

5.1 Germany

Germany is the country with one of the highest population in Europe and the 20th in the world with 81.2 million inhabitants at the end of 2014. It is located in central Europe. Altitude ranges from the Alpine mountain region to the beaches in the North and Baltic Sea. The four largest cities in the country are Berlin, Hamburg, Munich, and Cologne. The gross domestic product in Germany was 3.14 billion in 2016. Moreover, it has one of the lowest unemployment rates in Europe. Around 10% of the population is foreign ([Statistisches Bundesamt, 2016](#)).

Regarding the transport sector, Germany has a length of 12,900 km of roads. In 2013, the modal split consisted of the 22% walking, 13% cycling, 52% private car, and 13% public transport. The trips purposes are dived as 3.8% education, 12.3% work, 18.3% leisure, 21.5% shopping and services and 44.1% home related. Between 18 and 25 years old, 77% has a driver license, but they have just 27% of the time access to a private car. Around 48.9% of the population live under 10km from the workplace ([Statistisches Bundesamt, 2016](#)).

5.2 Study Cities

The cities included in the study are the German cities of Hamburg, Frankfurt am Main, Stuttgart, Kassel, Darmstadt, Marburg (see Figure 5.1). In the following section a brief overview of each city is presented highlighting the location, population and transportation aspects.

– **Hamburg** – is the second largest city in Germany with 1,814,597 inhabitants ([HAMBURG.DE, 2017](#)). Hamburg is a city-state with a system of self-government. It is located in Northern Germany, and it is one of the busiest ports in Europe. According to the transport sector, the harbor area and the city are served by the German railway network, and the city has public transport system based on buses and metro network ([Encyclopaedia Britannica, b](#)).

– **Frankfurt am Main** – is located in western Germany in the state of Hessen and it is crossed by the river Main. In 2016, 729,564 inhabitants lived in this city ([Stadt Frankfurt am Main, 2017](#)). This city is well known around the world because of its famous international trade fairs.

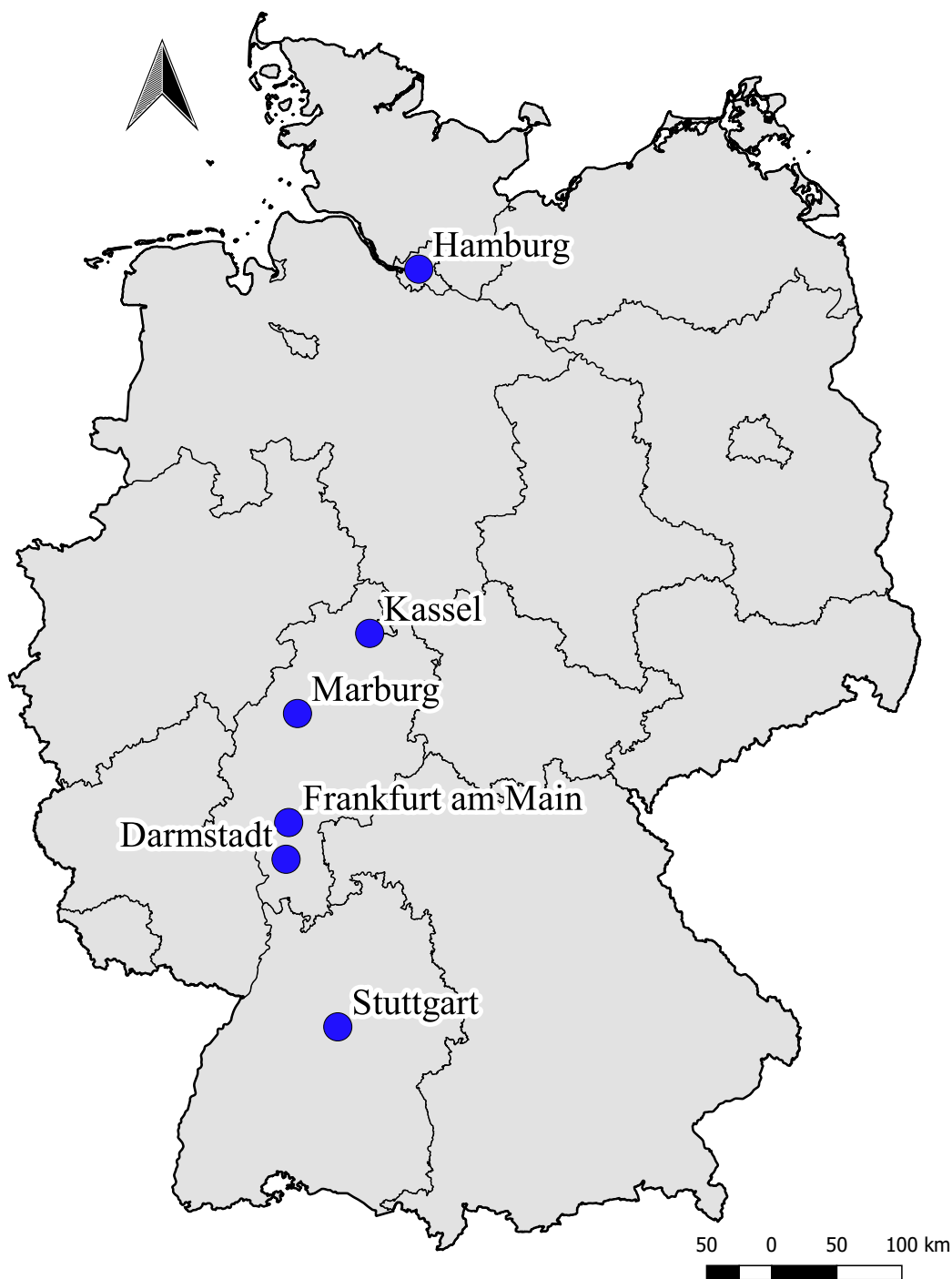


Figure 5.1: Location of the cities of the study

The Goethe University of Frankfurt is among the largest universities in Germany. The location of the city has allowed not only to be a strategic node for fluvial, rail and road transport but also it has the largest airport in Germany ([Encyclopaedia Britannica, a](#)).

– **Stuttgart** – is the capital of the state of Baden Württemberg in southwest Germany between the black forest and the Swabian Alps. It has a population of 623,738 inhabitants([Statista,](#)

2016). The city has a fluvial port since it is a node connected by two of the biggest rivers in Germany, the Danube river, and the Rhine River. It is the largest industrial zone of southwest Germany including the prominent companies of Daimler and Porsche. Its transportation systems are based on the regional train, buses, and trams. In the city, a polemic project Stuttgart 21 is ongoing that deals with the construction of an underground train main station to increase the tram network ([Encyclopaedia Britannica, c](#)).

– **Kassel** – is situated in the center of Germany in the north of the federal state of Hessen with a population of 201,907 inhabitants. It is called the "Documenta city" after the international art exhibition Documenta realized every five years. The university founded in 1970 was an important factor for its urban development. Because of its privileged location it is an important node in the railway network. The airport Kassel-Caden is located at 30 min from the city. Public transport in Kassel is well developed with a tram and bus network in the city and the surrounding area ([Stadt Kassel, 2017](#)).

– **Darmstadt** – is located in the federal state of Hessen, in the metropolitan region of Rhein-Main. It has a population of around 159,470 inhabitants. It was the capital of Hessen until 1945 because of its massive destruction during the II World War. Because of its scientific and cultural importance, it was awarded as a scientific city. In Darmstadt, tram and buses are the main public transport mode with frequencies of 15 minutes ([Wissenschaftsstadt Darmstadt, 2017](#)).

– **Marburg** – is a university city with around 72,000 inhabitants. It has a favorable location in the middle of the federal state of Hessen, around 30km south from Frankfurt am Main. It is an advantage for many business connections: one hour driving to Frankfurt am Main, and proximity to the highway network. The Phillips-university made Marburg a city that is internationally well-known with 26,000 students. Around 10,000 employees work at the university and the clinic of the university ([Universitätsstadt Marburg, 2017](#)).

5.3 Car sharing in Germany

Car sharing is available in Germany since 1988 after a project in Berlin called Stadt-Auto. Then, the first car sharing operators were born in 1990: STATTAUTO GmbH, and Stadt Auto Aachen and Bremen, today cambio. In 1991, the idea started to expand in whole Germany. The next years the operators in individual cities started to merge giving birth to AutoCarsharing in 1998 (since 2005 Greenwheels), Stadtmobil since 1999 and cambio since 2000. In 2009, car2go started the first free-floating car sharing service in Ulm by the company Daimler and in 2011 DriveNow in Berlin by BMW. From 2012 the other operators start the FFCS service as Stadtmodil under the name of stadtfliitzer and Book N Drive. Since March 2017 the German parliament recognized car sharing under the law ([CarSharing eV Bundesverband, 2017a](#)).

Because of the high demand, the past years, car sharing in Germany has experimented a high growth especially in FFCS (see Figure 5.2). There are around 1.7 million users of car sharing in Germany at the beginning of 2017 ([CarSharing eV Bundesverband, 2017b](#)) of 4 FFCS operators with 7800 vehicles in 12 cities and 150 SBCS operators with 9400 vehicles in 597 cities. The greatest FFCS companies are car2go (Daimler), DriveNow (BMW), and Multicity; and the main SBCS are Flinkster, stadtmobil, Cambio and teilAuto (see Figure 5.3).

5.4 Bike sharing in Germany

In 2014, Germany had the fifth largest fleet in the world with around 12474 shared bicycles ([Meddin and DeMaio, 2015](#)). Currently, there are four BS systems in Germany: Call a bike,

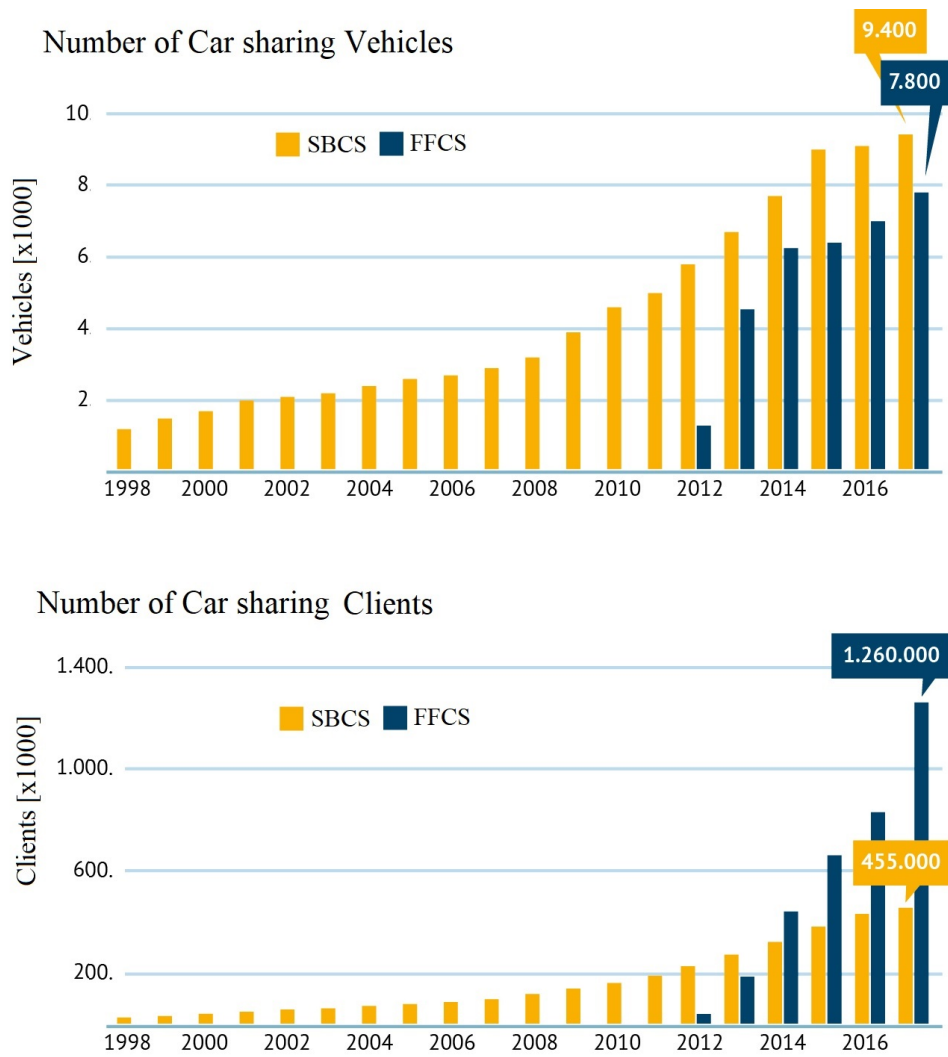


Figure 5.2: Evolution of car sharing in Germany (2016)

Source: CarSharing eV Bundesverband (2017b)

nextbike, Bikey and Chemnitzer Stadtfahrrad (Büttner and Petersen, 2011) and Obike. The market is dominated mainly by Call a bike and nextbike. Next bike is a company founded in 2004 in Leipzig. It has a fleet of around 20.000 bicycles (April 2015) in 15 countries in 4 continents. In Germany, Next bike works in 35 cities with free-floating or station based bike sharing systems (Deutsches Institut für Urbanistik, 2015). In late 2017, a new operator namely Obike supplied Munich with 7000 shared bicycles (two times more than the previous existing BS systems (Schubert, 2017)). In this thesis, the examined system is Call a bike (operated by DB). The main reason of focusing on this service is the open source dataset offered. The main characteristics of this system is presented below.

5.4.1 Call a bike

Call a bike is a bike sharing operator offered by the German train company Deutsche Bahn. It is a station based or free floating bike sharing system in around 50 cities in Germany. Their bicycles contain a small electronic box that regulates the lock. The design of the bicycles plays an important role because they are red and with unique physical characteristics that ease the users to identify them and also to recognize them in a theft situation. In Stuttgart and Aachen, Call a Bike offers electric bikes. When users are registered, they can use the system in all the

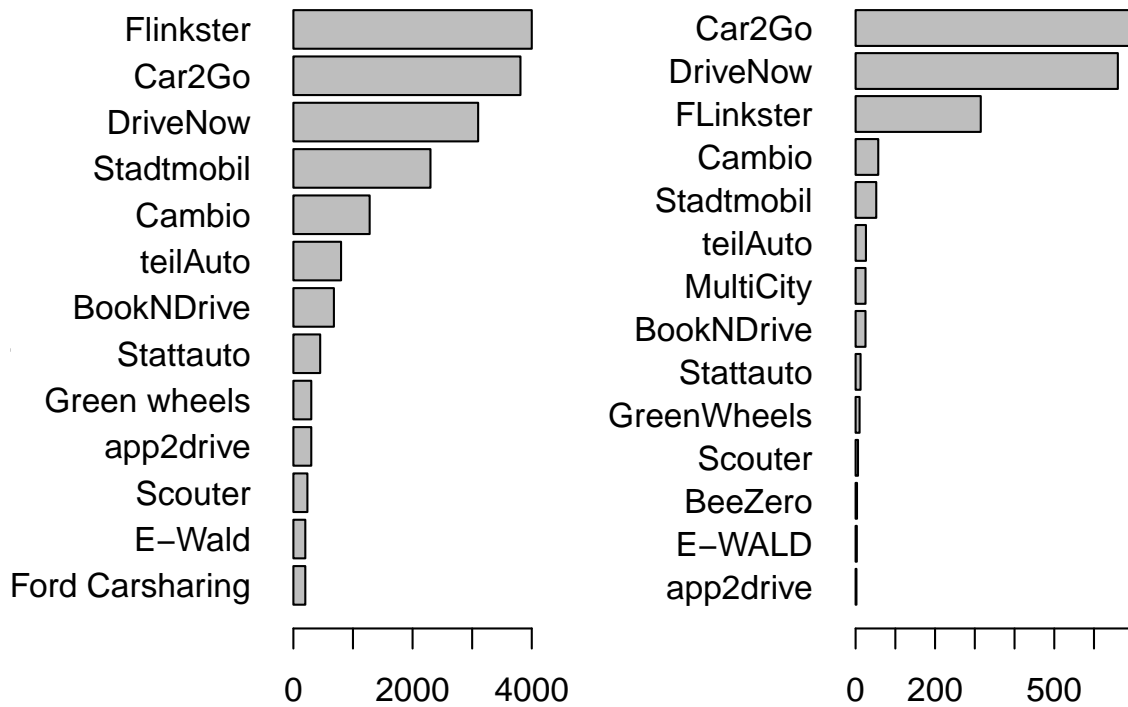


Figure 5.3: Fleet size of car sharing fleet size (left) and clients (right) in Germany (2016)

Source: [Carsharing News \(2017\)](#)

50 cities that count with Call a Bike, StadtRAD in Hamburg, StadtRAD in Lueneburg and Konrad in Kassel ([Deutsches Institut für Urbanistik, 2015](#)). In Munich and Cologne, Call a Bike offer a FFBS system [Deutsche Bahn AG \(2017\)](#).

To use a bicycle from Call a Bike, the user has to create an account. Then the rentals can be realized through the App, a telephone call, a client card or direct at the stations with interactive terminals. For the call, users have to deal the printed number on the electronic box. Then, they get a code to write on the box to free the lock. From this moment the timing starts. Pauses during the drive are allowed. To return the bicycle, users have to lock the bike to the station or to lock themselves in FFBS services. There are three types of registration tariffs: a) the basic of 3 EUR per year, b) the "comfort" of 49 EUR a year or 7 EUR per month and c) a day ticket for 15 EUR. The cost per minute is 0,08 EUR, for the comfort tariff the 30 first minutes are free. There are discounts for students and members of the German Train [Deutsche Bahn AG \(2017\)](#).

Call a Bike works in 50 cities, as mentioned before, however, this thesis studies only six of them: Hamburg, Frankfurt am Main, Kassel, Stuttgart, Darmstadt, and Marburg. Even Call a bike is one of the BS operators in these cities, it works differently in each of them as presented below.

– **Hamburg** – It is the city with the highest demand of Calla Bike in Germany. However, in this city takes the name of "StadtRAD". It is financed entirely by the city and it has a high acceptance between its inhabitants ([Deutsches Institut für Urbanistik, 2015](#)). This system started in July 2009 with 69 stations. Nowadays, it has 2450 bicycles traveling around 206 stations. The first 30 minutes of the rental are free, and from the 31st minute, the cost is 0.08 EUR with a maximum of 12 EUR per day. It is one of the most used BS systems in Germany with

360.000 memberships, where 2% use them every day and 50% combines regularly StadtRAD with the public transport. According to the trips, Figure 5.4 shows the share of different trip purposes for StadtRAD. The most common trip purpose was leisure activities with the 56% of share. 60% of the trips are less than 15 minutes and 29% between 16 and 30 minutes. The peak season is from June to August and the off-peak season is from December to February. During the workdays, the morning peak hour is at 8:00 and the afternoon peak hour is at 18:00. Weekends have a different behavior, with one peak period from 13:00 to 18:00 (Böhm, 2016).

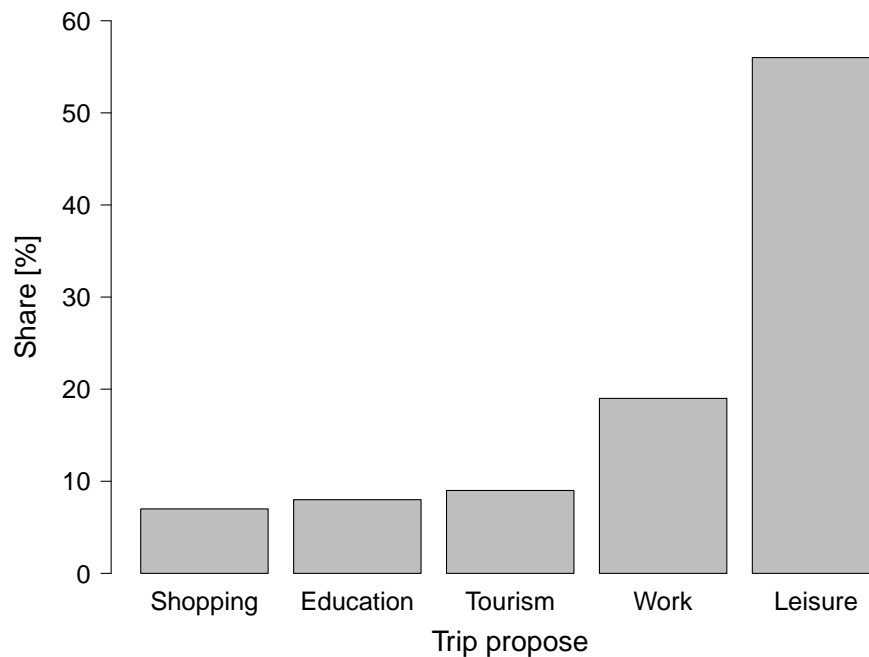


Figure 5.4: Trip purposes for StadtRAD users (2016)

Source: Böhm (2016)

– **Frankfurt am Main** – This city has the second biggest demand of Call a Bike in Germany. It started in 2004 and now it counts with 2300 bicycles working on 350 stations in the whole metropolitan area and 318 inside the city (DB Rent GmbH , 2015).

– **Kassel** – This city has an innovative BS system called "Konrad" that integrates the BS with public transport. From 2014, Konrad belongs to the German train company and therefore, any Call a Bike user in Germany is allowed to use Konrad without a new registration (Stadt Kassel, 2016). It is founded by the city and supported by the German train company (Deutsches Institut für Urbanistik, 2015). Since March 2012, 500 bicycles are available within 58 stations.

– **Stuttgart** – This city counts with the system of Call a Bike with electric bicycles (pedelecs) since October 2011. 44 stations, 400 bicycles, and 100 pedelecs facilitate biking through the inconvenient topography and also for older users. Regular bikes have a cost of 0,08 EUR per minute (15 EUR per day) and pedelecs a fee of 0.12 EUR per minute (22.5 EUR per day) (Deutsches Institut für Urbanistik, 2011).

– **Darmstadt and Marburg** – The students association from Darmstadt, Marburg, and Main started the project to bring Call a Bike to their cities. Students pay an extra fee to the

universities of 2,38 EUR per semester to contribute to the BS systems. Thus, students are allowed to drive the shared bicycles for one hour without extra cost. From the 61st minute they have to pay a fee of 0.08 EUR per minute. Not students have to pay annually 39 EUR or monthly 7 EUR to get the same benefits. Marburg has available 200 bicycles with 22 stations ([Gießener Anzeiger Verlags GmbH, 2017](#)). From April 2014, Call a bike operates in Darmstadt with more than 350 bicycles in with 41 stations.

Chapter 6

Results

This section reports the data collection, analyze and process. The data collected were the rentals and stations of six cities in Germany from the SBBS "Call a Bike." Existing infrastructure data was collected from Open Street Map. This dataset was analyzed using several plots and correlation matrices. Finally, time intervals, zones of influence and a pre-selection of the independent variables were set.

6.1 Data collection, analysis, and processing

6.1.1 Data collection

This research focused on the German bike sharing system "Call a Bike" Deutsche Bahn (DB) (2017). Arrivals and departures of bicycles were downloaded from the Open-Data-Portal offered by the German train company (Deutsche Bahn) under the link: <http://data.deutschebahn.com/dataset/data-call-a-bike>. The dataset (around 3.96 GB) included the rentals in fifty cities in Germany in approximately 3.5 years (from 01-2014 to 05-2017).

Six station-based bike sharing systems in six cities in Germany were selected to build the models: Hamburg, Frankfurt am Main, Stuttgart, Kassel, Darmstadt, and Marburg. These cities were selected based on their high usage of bike sharing (>250,000 rentals) (see Figure 6.1). This threshold value was set because of the significant step between rentals in Marburg and Rüsselheim (see Figure 6.1). Of the selected cities, Munich and Cologne are free-floating-based systems, and their arrivals and departures were aggregated into symbolic stations. However, they are prone to large topological errors (e.g. they were only located in some parts of the city and not spread in the whole business area), therefore these cities were removed from the study. Berlin was excluded because of the lack of location data of the stations.

Within these six cities, three types of open data were collected:

1. Rentals (Time and station's ID of the arrivals and departures).
2. Stations (location)
3. Exogenous factors

– **Rentals** – The selected cities represented 91% of the total 13.13 million rentals included in the original dataset. Each rental data contained the following information:

- ID (e.g. 21366843)
- Vehicle ID (e.g.143517)
- Departure date and hour (e.g. 2014-01-01 00:34:54)
- Arrival date and hour (e.g. 2014-01-01 00:50:14)

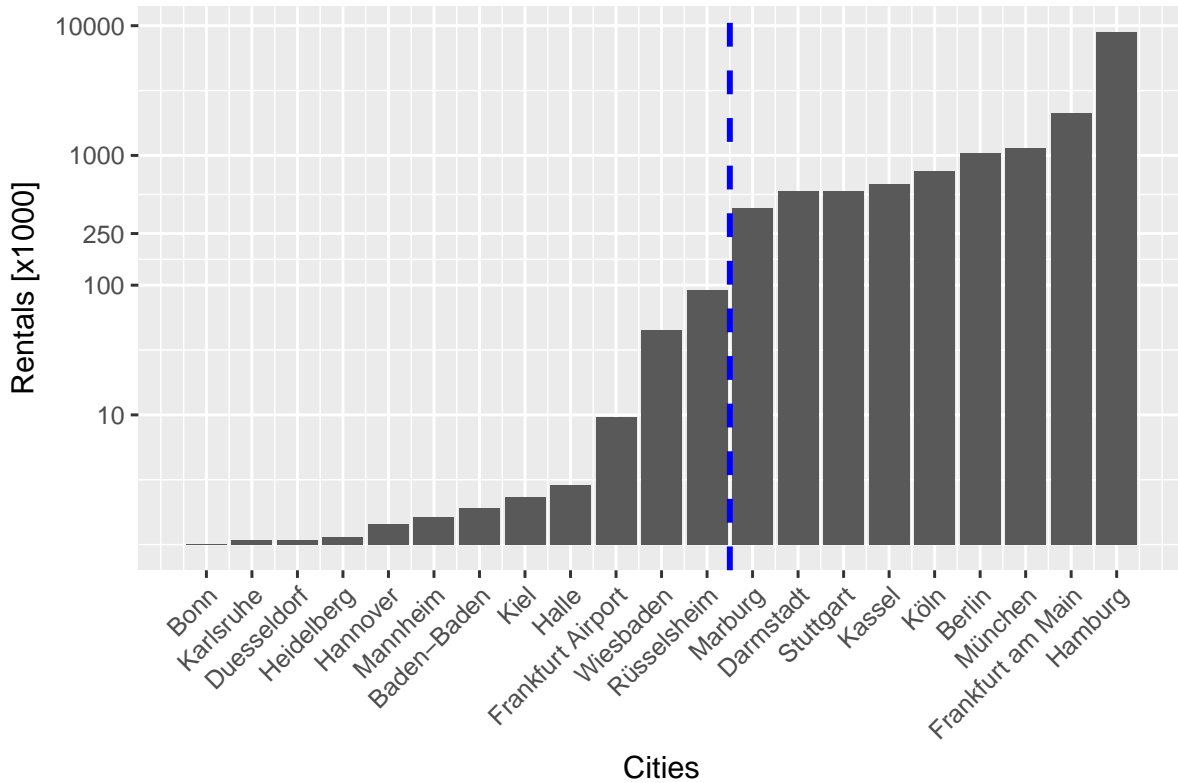


Figure 6.1: Call a Bike: Rentals per city (01/14 - 05/2017)

- Departure station ID (e.g. 214170)
- Arrival station ID (e.g. 131880)
- City (e.g. Hamburg)

– **Stations** – The "rental zones" in the Open-Data-Portal correspond to the fixed stations of Call a Bike. Figure 6.2 shows a map where the location of the stations in the six cities of the study. Each station provided the following information:

- ID (e.g. 250380)
- Name (e.g. Melanchthonplatz)
- Latitude (e.g. 8.727924)
- Longitude (e.g. 50.098068)
- City (e.g. Frankfurt am Main)
- Country (e.g. Deutschland)

– **Exogenous factors** – Three types of exogenous factors were analyzed:

1. **City size.** The city size is the equivalent of the total city population, which was collected from the last census [Statistisches Bundesamt \(2012\)](#).
2. **Population density** The population density data were obtained from [Suche-postleitzahl.org \(2017\)](#) for each postal code area in Germany.

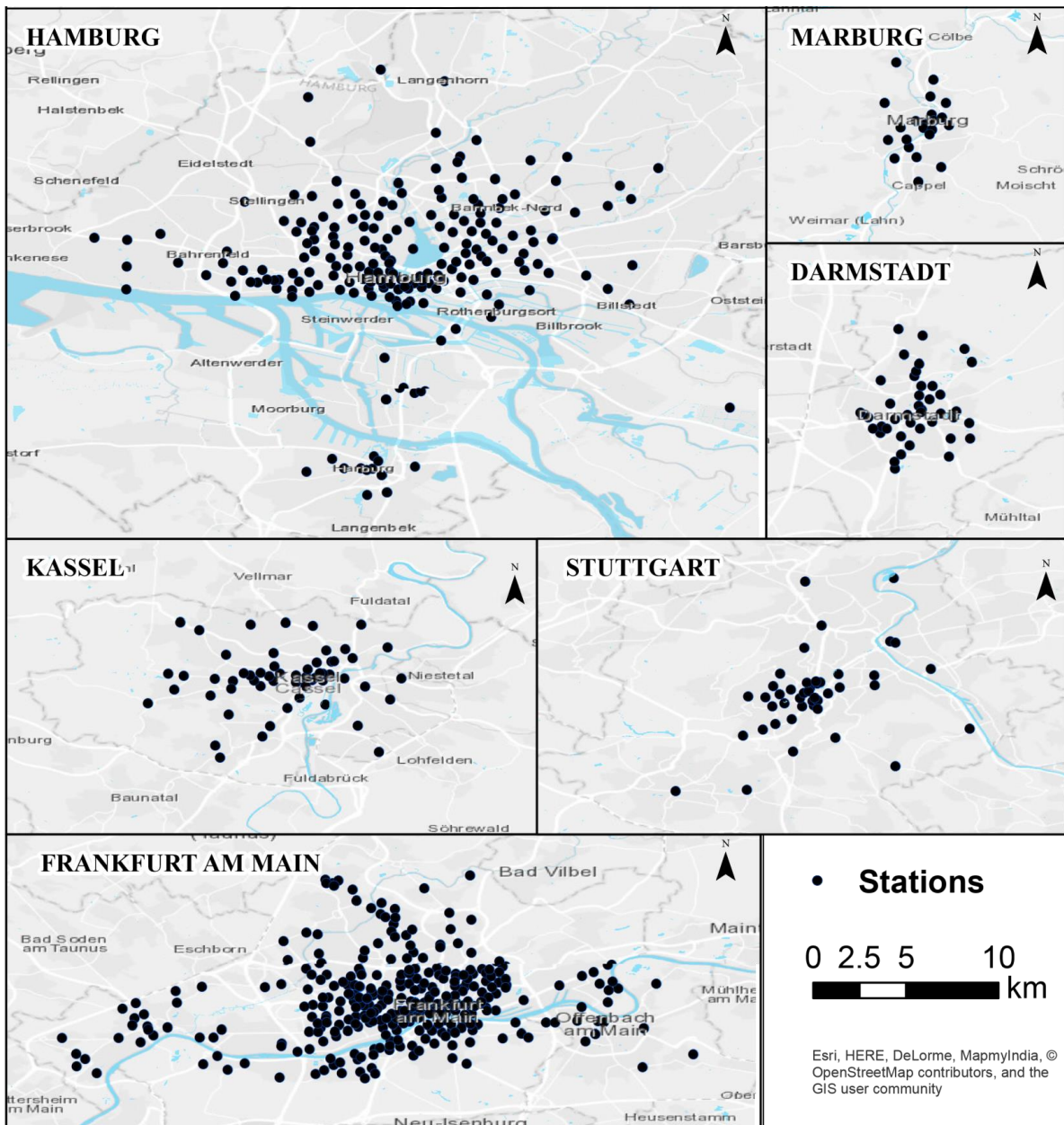


Figure 6.2: Location of the stations

3. **Existing infrastructure.** The existing infrastructure was downloaded for the German Federal states of Hamburg, Hessen, and Baden Wurttemberg through the website: Geofabrik GmbH Karlsruhe (www.geofabrik.de/data/download.html), which is a member of the [OpenStreetMap-contributors \(2017\)](#). The downloaded information was categorized based on their spatial shape:

- *Points*: natural features, administrative areas, places of worship, points of interest, traffic-related, transport-related.
- *Lines*: railways, roadways, waterways.
- *Polygons*: buildings, land-use, natural features, administrative areas, places of worship, points of interest, traffic-related, transport-related, water bodies.

Furthermore, each category was subdivided into feature classes. For instance, the category "points of interest" has the feature classes bakeries, banks, hotels, restaurants, supermar-

kets, among others. The distance to the city center was also considered in the existing infrastructure variables since it is a factor considered in the literature Table 2.5.

6.1.2 Exploratory data analysis (I)

– **Dependent variables** – In total, 1.05 million rentals were included for the research. They took place from the 01.01.2014 until the 15.05.2017 (1232 days). However, the cities of Darmstadt and Marburg, which initiated the operation later, had just 1150 and 1141 days of service respectively. Around 73% of the rentals belong to the city of Hamburg (see Figure 6.3), followed by Frankfurt with the 12%. After Frankfurt, the following cities did not present a significant demand difference between each other. The considerable amount of rentals represent the success of the SBBS system in Hamburg. Even though it has more than double the population, it presented around six times more rentals than Frankfurt, although Frankfurt has even a higher number of stations. To support this fact, Figure 6.4 indicates that Hamburg has the highest rate of trips per day per thousand inhabitants, but Frankfurt has one of the lowest usage rates of the selected cities. Moreover, according to this rate, the second most successful system is Kassel, followed by Marburg.

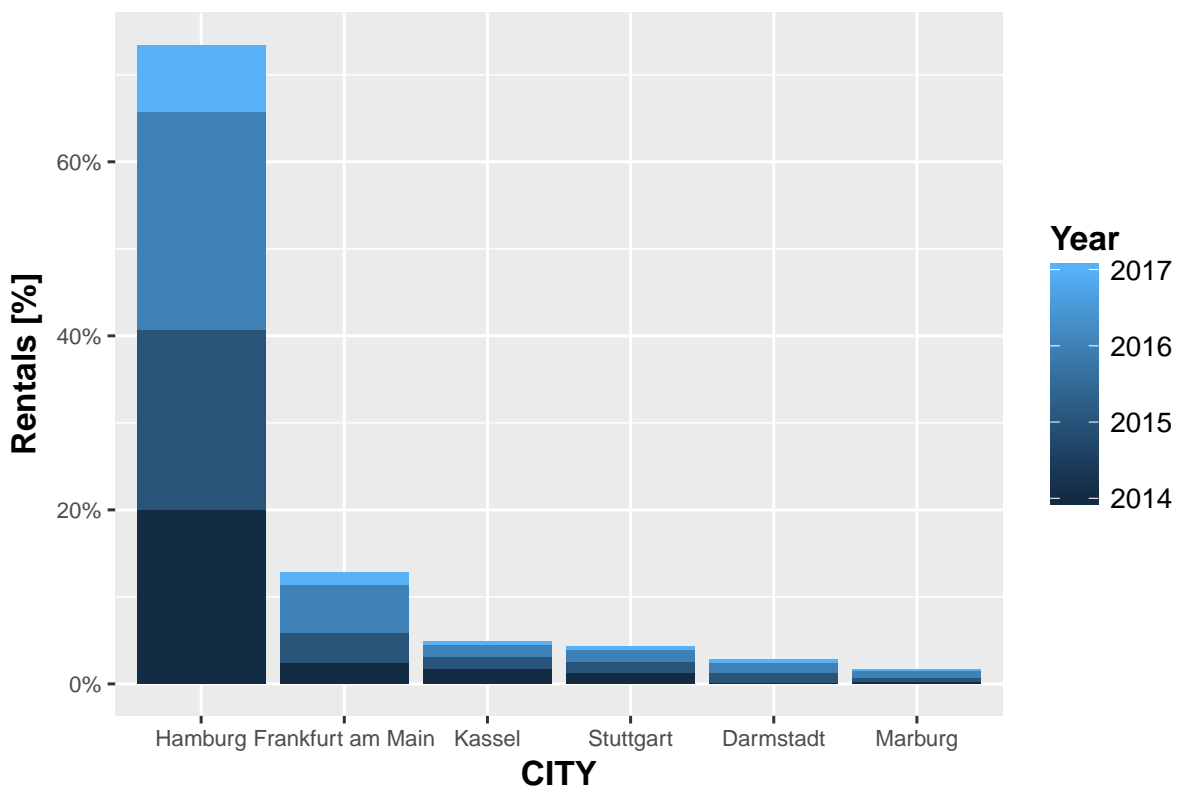


Figure 6.3: Distribution and growth of the rentals in the cities of the study

In average, 9781 trips per day are made in the six cities. Hamburg has the highest rate with a mean of 7184 trips per day. There is a notable change to Frankfurt, with 1256 trips per day. The system with the least frequency was Marburg with 176 trips per day (see Figure 6.5). Although some outliers were present, we did not remove them. The two main reasons were: 1) because they are real data and 2) because of the high quantity of data, they will not affect the average significantly. Outliers were possibly due to massive events such as concerts or parades.

The travel time and the travel distance were also analyzed. The median of travel time is 12.9 minutes, and the median of the travel distance is 1.62 km. In contrast to the daily trips, the average travel time and travel distance do not vary significantly from one city to another. The

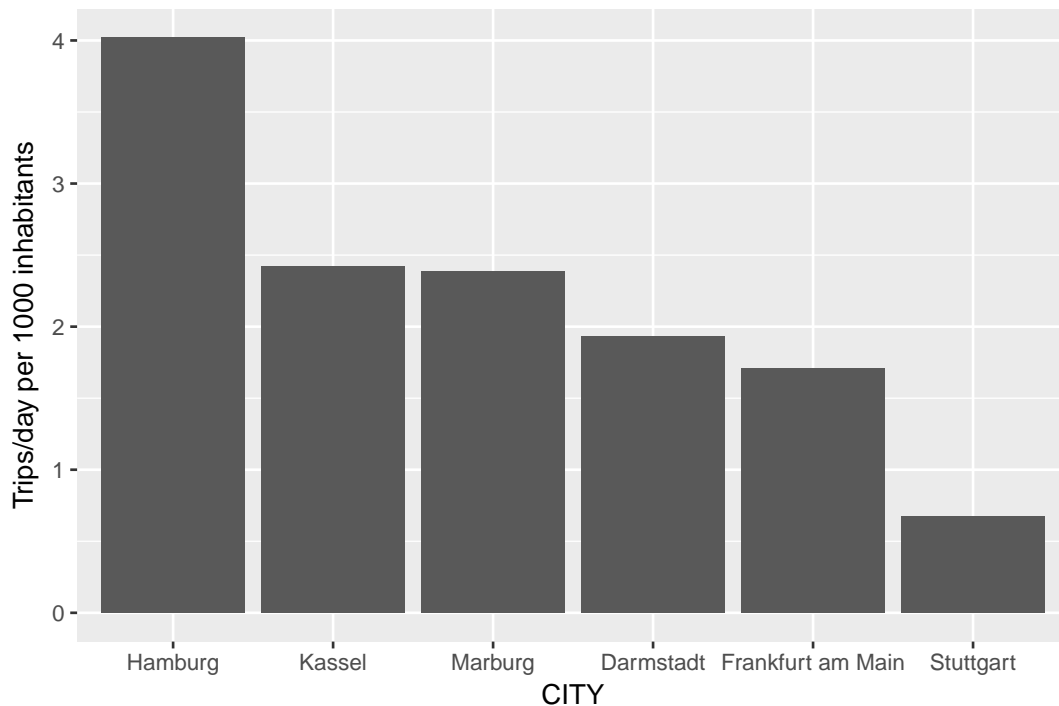


Figure 6.4: Trips/day per 1000 inhabitants

range of the median travel time varies from Hamburg, which presents a travel time median of 12.55 minutes, to Stuttgart, with a median of 8.36 min. In the case of the traveled distance, their medians change from 1.7 km in Hamburg to 1.21 km in Marburg (see Figure 6.5). Furthermore, higher travel times were estimated between 12:00 and 20:00, and shorter between 5:00 and 9:00.

Four cities of the study showed an increasing average of daily trips over the considered period. The steepest growth occurred in Darmstadt and Marburg. However, Stuttgart displayed a relatively little growth, and Kassel is the only city that presented a decreasing trend of daily trips (see Figure 6.6).

Rentals varied significantly with season. Peaks were present in the summer time (May to July), and troughs in winter time (December to February) (see Figure 6.7). Usually, the cities show fairly similar behavior over each month, with differences seen mainly in March and June. In April, an increase of rentals was observed in smaller cities after the winter time but for bigger cities this increase was in March. The bigger cities decreased their usage in June in relation with May and July. The months with the most common behavior between cities were April and December.

Looking at the daily data, Wednesdays and Thursdays are the days that showed the highest demand. The demand decreased on the weekends (see Figure 6.8). Friday presented the least spread between cities of all the days, whereas Saturdays and Sundays showed more spread, having lower demand in Darmstadt and Frankfurt in comparison to Hamburg and Kassel. Regarding the hourly behavior, there is a different trend between workdays and weekends (Fig 6.9). A more steady hourly change is on the weekends with a peak period from 13:00 to 19:00 and an off-peak hour at 6:00. Also, there is a relatively small peak at 24:00. In Darmstadt, the behavior is more steady in contrast with Hamburg that showed a higher demand in the afternoon.

During workdays, the demand changed between cities, especially at 8:00 and at 18:00. The off-peak hour was at 5:00 and the peak-hour differed from city to city. However, we can see that there are two peak periods at 8:00 and at 17:00 (based on the median values). In Frankfurt, the demand was much higher than the other cities at 8:00, and in Darmstadt, it was relatively higher at 13:00. To conclude the analysis of the dependent variables, Figure 6.10 shows the

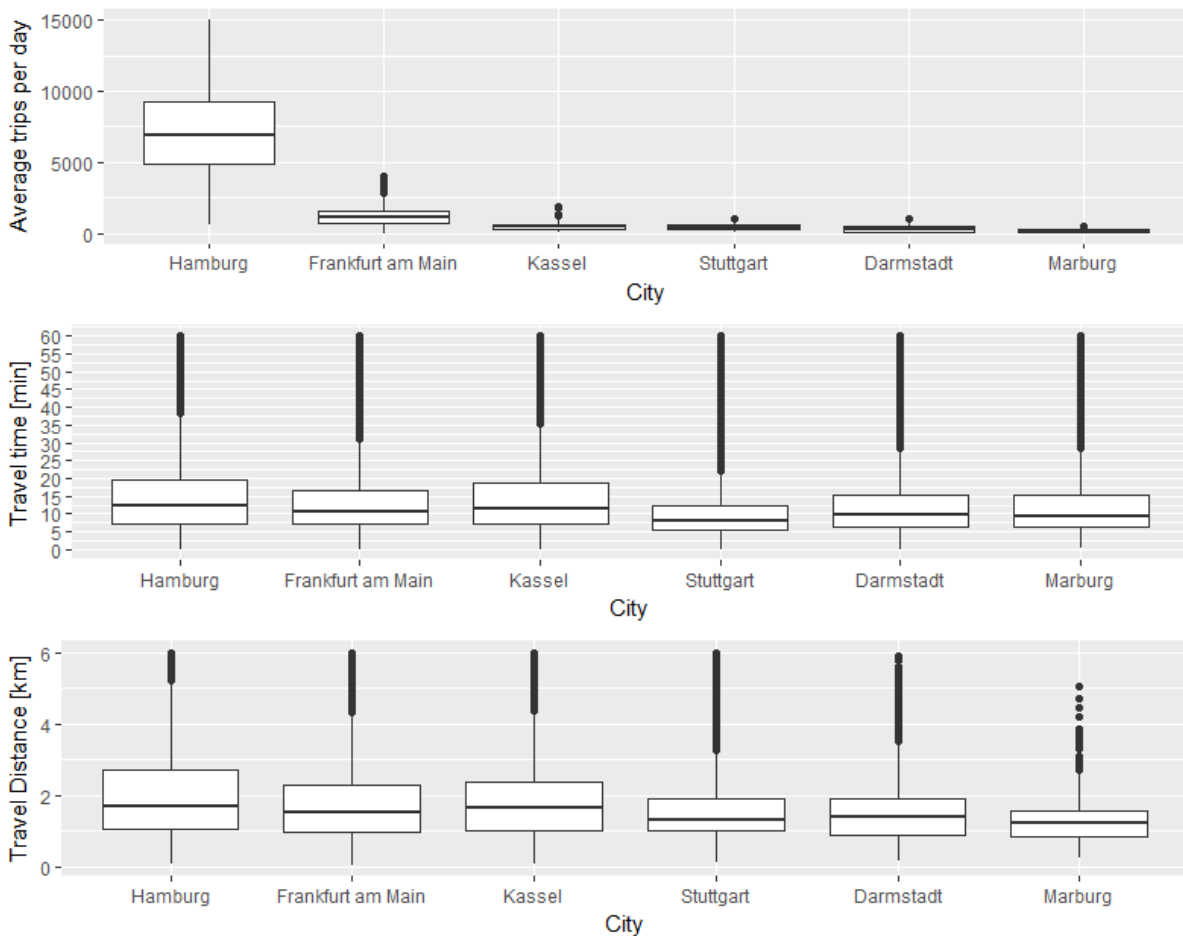


Figure 6.5: Analysis of trips performed

spatial distribution of the intensity of the rentals. Each area represents the frequency of a station with the help of Voronoi Diagrams for a better visualization. These areas are not the zones of influence. As it is evidenced in most cases, the number of rentals are higher in areas near to the city center. However, the spatial distribution of frequencies was more uniform in smaller cities like Kassel, Marburg, and Darmstadt,

– **Independent variables** – The spatial independent variables were obtained from OpenStreetMap ([OpenStreetMap-contributors, 2017](#)). [Haklay \(2010\)](#) and [Zielstra and Zipf \(2010\)](#) studied the quality of Open Street Maps (OSM) data presented in England and Germany. A high accuracy of six meters and an acceptable overlap of roads were present in England. However, there was a clear difference between the completeness of some areas, mainly urban vs. rural. [Haklay \(2010\)](#) concluded that OSM data in England have an expected accuracy of 70%, with an occasional reduction to 20%. This statement is not much different than commercial datasets [Haklay \(2010\)](#). Similar results were extracted for OSM in Germany. In areas as Hamburg, Frankfurt and Stuttgart OSM are even better than TeleAtlas (a company that counts with GPS and laser scanner with 1m accuracy ([www.teleatlas.com](#)) in the street network set ([Zielstra and Zipf, 2010](#)). Moreover, street data quality is satisfactory in major cities in Germany and still acceptable in midsize towns. Therefore, in the context of the thesis, this information is relevant especially given the fact that this study focuses on an analysis of the deployment of bike sharing systems in urban areas. Then, we can be sure that this data have an acceptable accuracy to meet the objectives of the thesis.

The main categories and features obtained from the OSM dataset were:

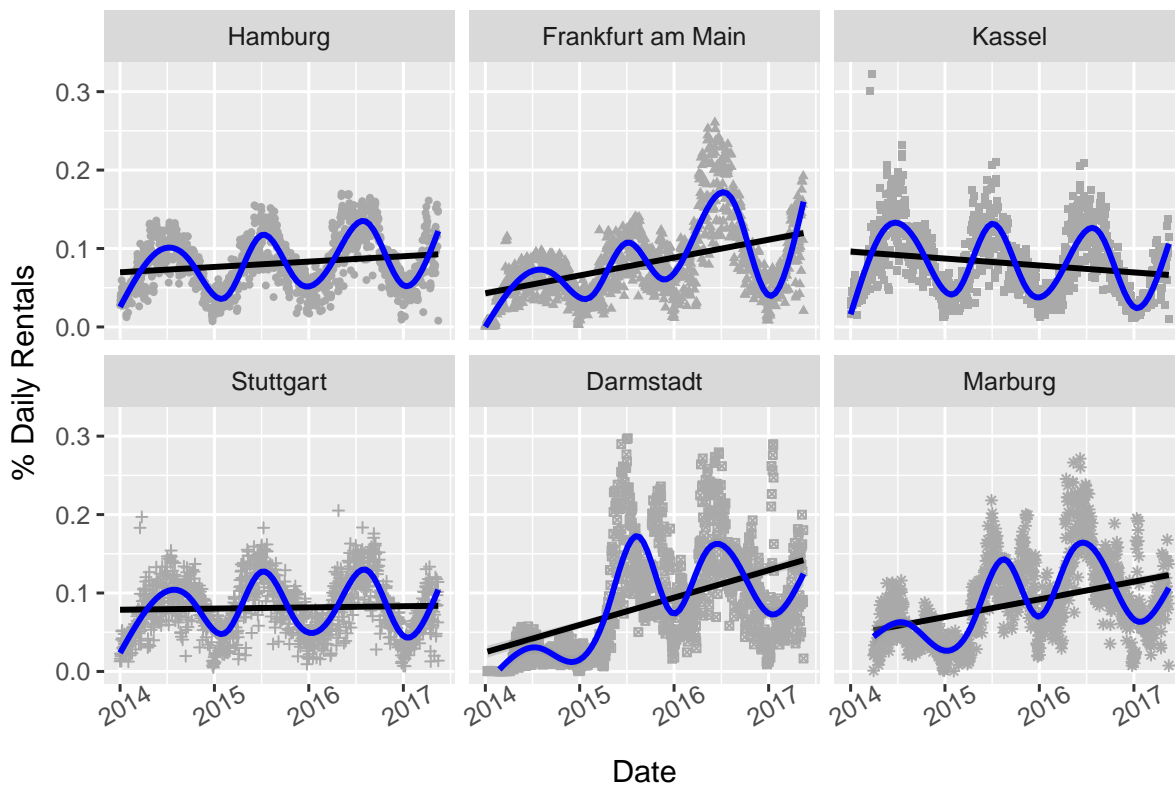


Figure 6.6: Monthly rentals vs month

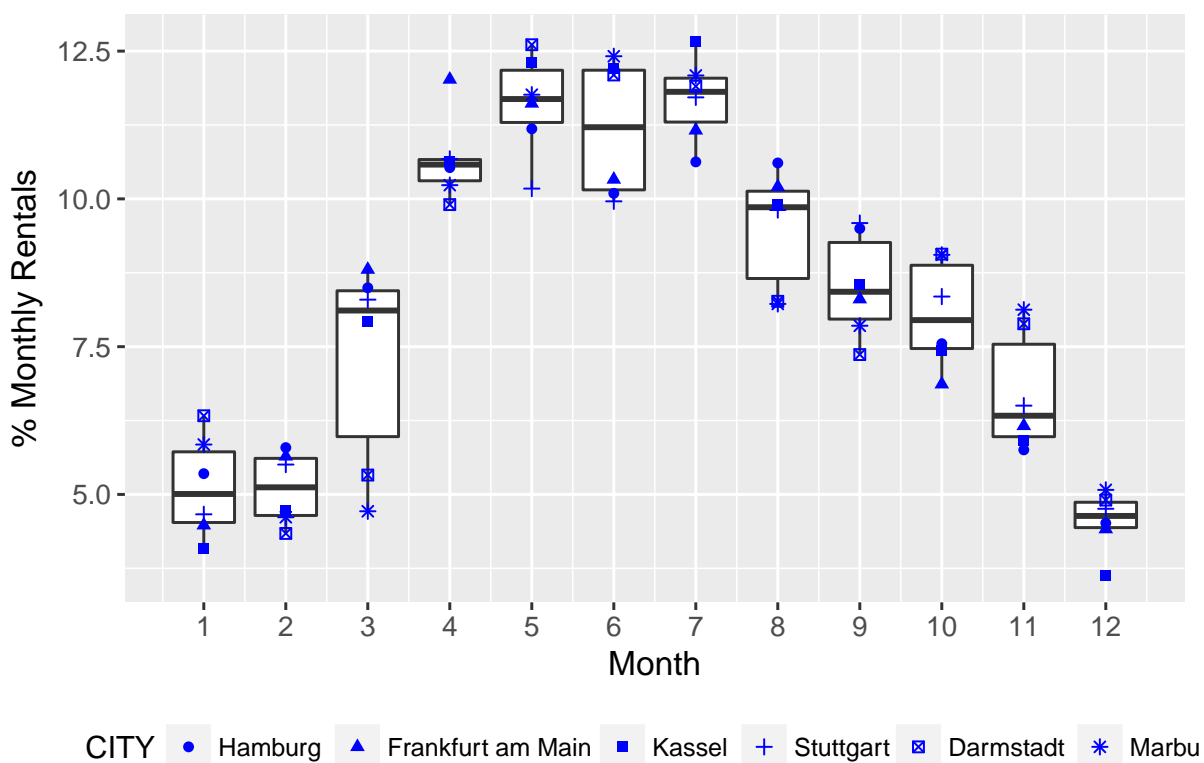


Figure 6.7: Monthly rentals vs month

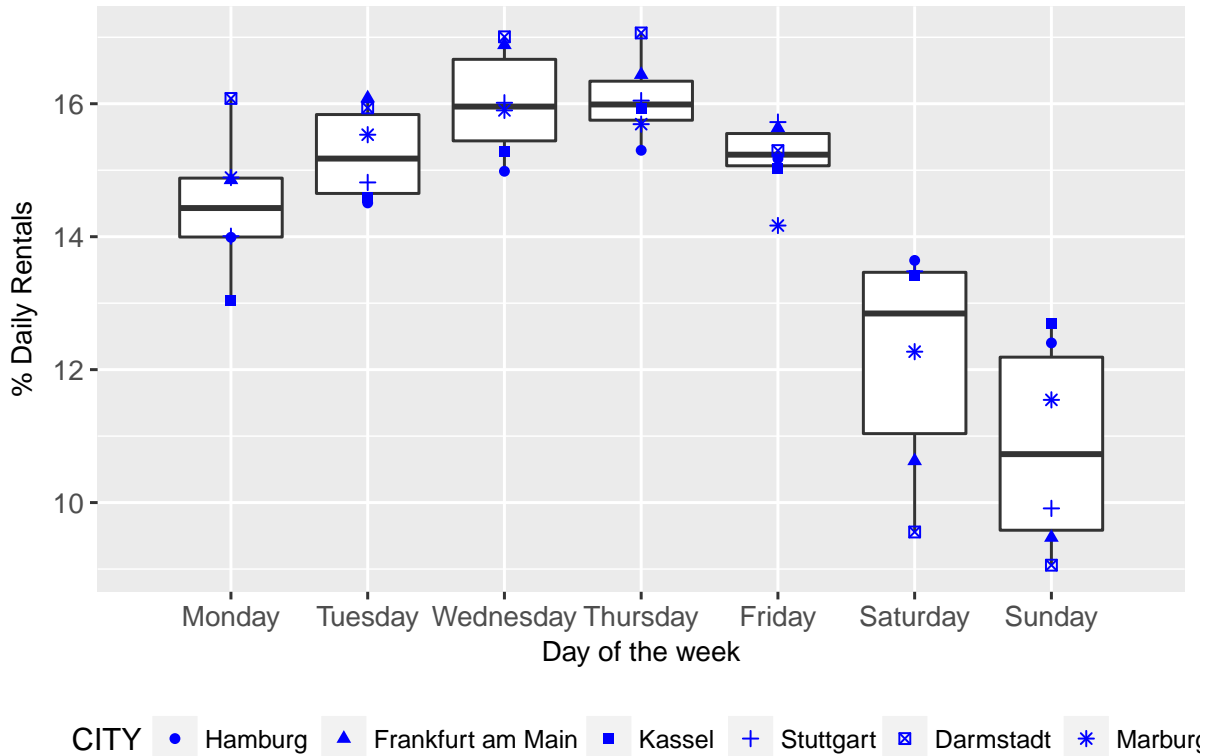


Figure 6.8: Daily rentals vs. day of the week

1. Points:

Points of interest: bakeries, banks, hotels, restaurants, supermarkets, etc.

Traffic-related: pedestrian crossings, traffic signals, etc.

Transport-related: public transport stops and taxi stops

Natural features: trees, springs, etc.

2. Lines:

Railways

Roadways: cycle-ways, foot-ways, residential streets, track streets, highways, etc.

Waterways: rivers, streams, canals, etc.

3. Polygons:

Land-use: commercial, forest, parks, residential, etc.

Natural features: cliffs, beaches, riverbanks, etc.

Traffic-related: parkings, gas stations, etc.

Water bodies: rivers, lakes, etc.

Not all the features were considered for the research. Some were removed taken into account the criteria from Section 3.1. Thus, some variables were not studied because of the following reasons:

- The polygons of the buildings were not selected for the study. The names of the buildings categories were set by the contributors of OSM, so there were incorrect or not existing features classes.

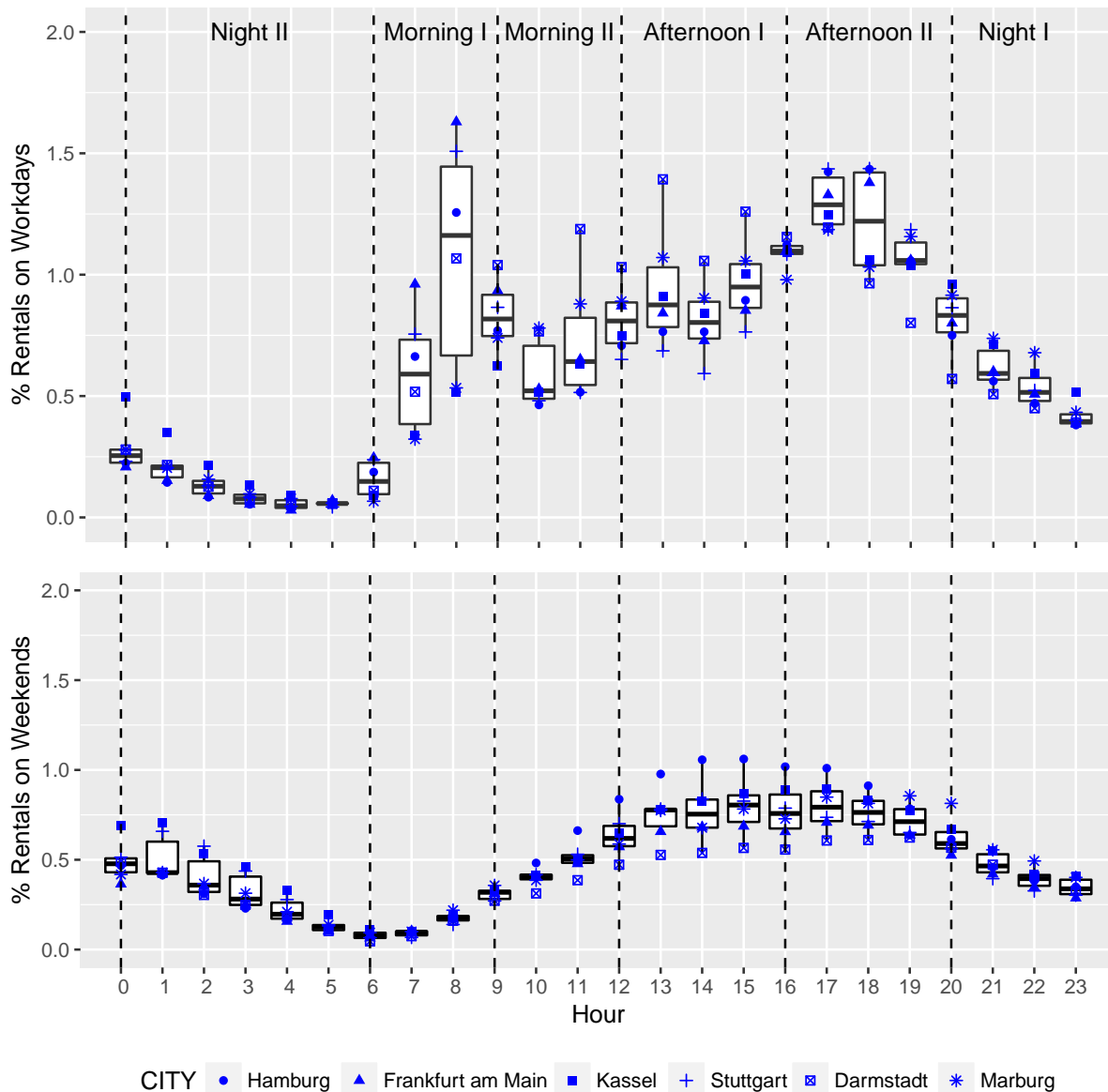


Figure 6.9: Hourly distribution and definition of times intervals

- Unclassified roads were obviated because of the lack of accuracy.
- Because of the repetitiveness of features classes, the polygons corresponding to the points of interest were not taken into account. The reason is due to the better accuracy of the points than the areas of the POIs.
- The categories of places (administrative divisions) was not considered because of the lack of relevance for the research. Moreover, the features classes considered as not relevant for the study and consequently, removed were: vending machines, wastebaskets, telephone boxes, post boxes, atm's, recycling centers for clothes or glass, public toilets, benches, among others.
- Features track, track grade 1, 2, 3, 4 and 5 were aggregated as one variable called "track all."
- Bicycle rental stations were not selected because they might influence the results of the model.

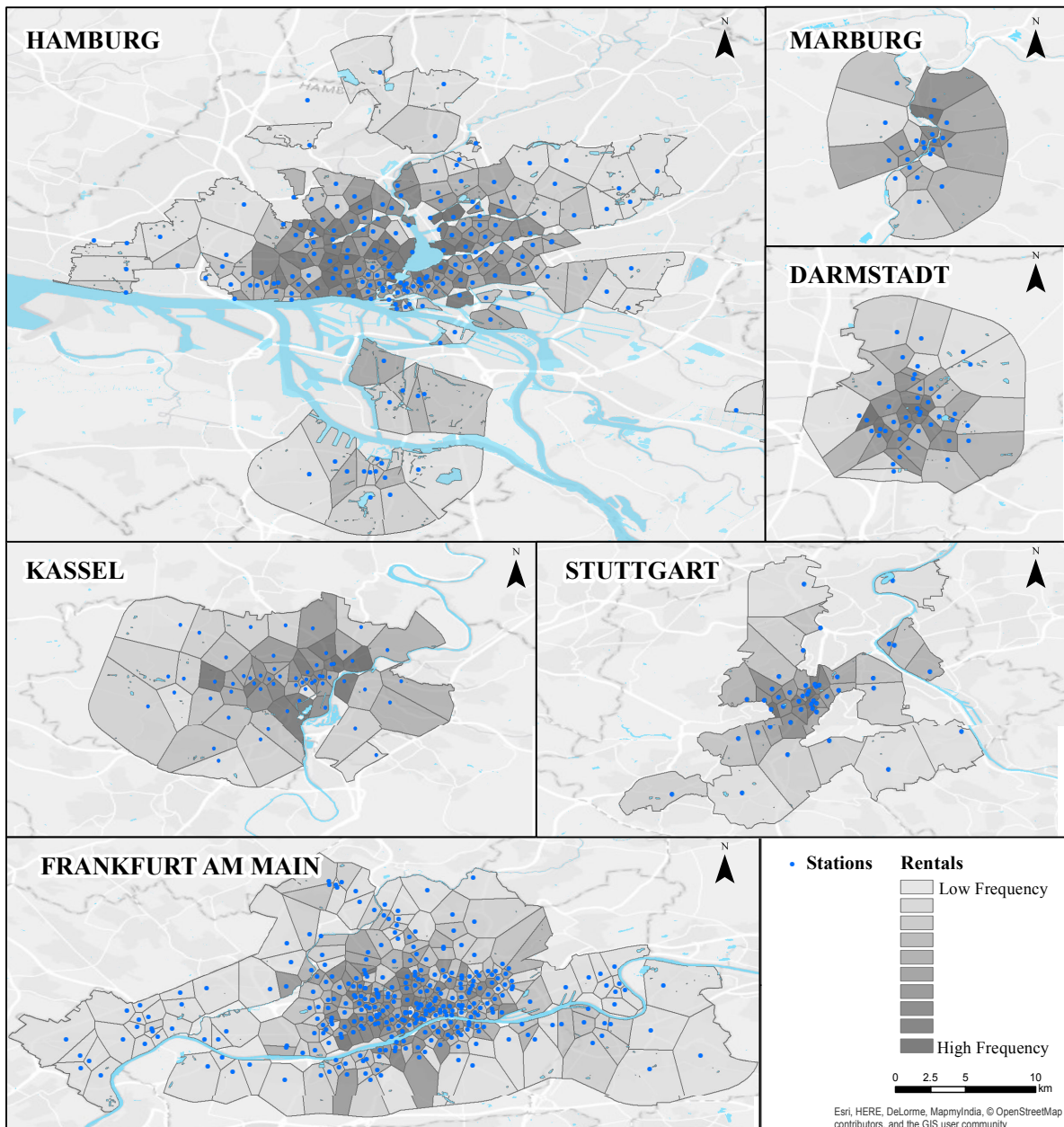


Figure 6.10: Spatial demand distribution in cities of the study

6.1.3 Time intervals

The goal of this thesis is to derive models that are as precise as possible. Thus, arrivals and departures were aggregated into day intervals. Day intervals were defined according to the hourly rental patterns. They represent peak and off-peak periods at the morning, afternoon and night: Night II (0:00-6:00), Morning I (6:00-9:00), Morning II (9:00-12:00), Afternoon I (12:00-16:00), Afternoon II (16:00-20:00) and Night I (20:00-24:00)(see Figure 6.9).

After each day of the week might have a different behavior, we classified the data also into days of the week. To reduce the number of output models, we can cluster the days of the week with similar behavior. Therefore, a Pearson correlation analysis was carried out to determine the most correlated days of the week. On Figure 6.11, we can see that weekends were less correlated with workdays at "Morning I", "Morning II", "Afternoon I" and "Afternoon II". Even between Saturdays and Sundays, there was not a relatively high correlation. "Night I"

presented a slightly different behavior on Fridays and Saturdays about the other days of the week. Finally, "Night II" is a special case where all the days of the week have a relatively high correlation.

For simplicity and considering the small differences on "Night I" and "Night II", the days of the weeks are a random Workday (e.g., Monday), Saturday and Sunday. The average of workdays is not considered because they might influence the precision of the results. Naturally, the optimal case would be to have the seven days of the week, but since they are very correlated, we take just one of them.

In conclusion, we aggregated 18 time units: 6 day intervals times 3 types of days. Moreover, we considered attractions (arrivals) and productions (departures). So, we will have 36 time units as the outcome, which means 36 sets of dependent variables, in other words, 36 different models. For simplicity, we abbreviated each name of each time unit for further plots and tables using the nomenclature shown in Table 6.1.

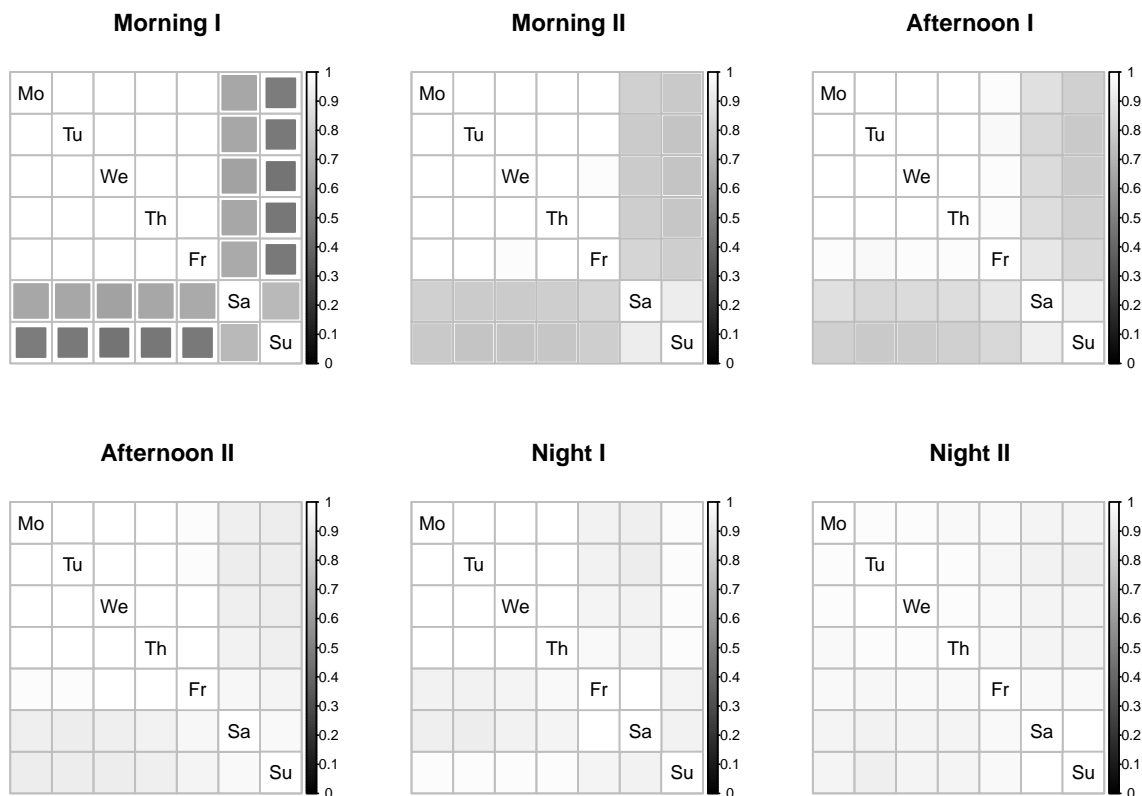


Figure 6.11: Correlation matrices between days of the week by time interval

6.1.4 Zones of influence

As discussed in the methodology, to build the zones of influence we intersected the postal code zones, the riverbanks, the voronoi diagrams and the buffer area from the stations. The postal code zones from Germany were obtained from Suche-postleitzahl.org (2017) and the riverbanks from the OSM dataset. For the buffer radius, a sensitivity analysis was accomplished (see Table 6.2) with the two most common values on literature: 300 meters and 400 meters (see Table 2.6).

The comparison criteria to select the buffer radius were the total number of variables, the number of the non-collinear variables, the average of the R^2 and R^2 adjusted of the 36 models,

Table 6.1: Nomenclature for the different dependent variables

	Dependent variables	Code
1	Monday Afternoon I Production	WA1p
2	Monday Afternoon II Production	WA2p
3	Monday Morning I Production	WM1p
4	Monday Morning II Production	WM2p
5	Monday Night I Production	WN1p
6	Monday Night II Production	WN2p
7	Saturday Afternoon I Production	SaA1p
8	Saturday Afternoon II Production	SaA2p
9	Saturday Morning I Production	SaM1p
10	Saturday Morning II Production	SaM2p
11	Saturday Night I Production	SaN1p
12	Saturday Night II Production	SaN2p
13	Sunday Afternoon I Production	SuA1p
14	Sunday Afternoon II Production	SuA2p
15	Sunday Morning I Production	SuM1p
16	Sunday Morning II Production	SuM2p
17	Sunday Night I Production	SuN1p
18	Sunday Night II Production	SuN2p
19	Monday Afternoon I Attraction	WA1a
20	Monday Afternoon II Attraction	WA2a
21	Monday Morning I Attraction	WM1a
22	Monday Morning II Attraction	WM2a
23	Monday Night I Attraction	WN1a
24	Monday Night II Attraction	WN2a
25	Saturday Afternoon I Attraction	SaA1a
26	Saturday Afternoon II Attraction	SaA2a
27	Saturday Morning I Attraction	SaM1a
28	Saturday Morning II Attraction	SaM2a
29	Saturday Night I Attraction	SaN1a
30	Saturday Night II Attraction	SaN2a
31	Sunday Afternoon I Attraction	SuA1a
32	Sunday Afternoon II Attraction	SuA2a
33	Sunday Morning I Attraction	SuM1a
34	Sunday Morning II Attraction	SuM2a
35	Sunday Night I Attraction	SuN1a
36	Sunday Night II Attraction	SuN2a

and count of better models comparing with R^2 adjusted. Furthermore, the regression models implemented were a simple OLS and the stepwise regression using as dependent variable the original dataset and its logarithmic transformation. The logarithmic transformation was performed after its common use in the literature (see Table 2.4). As a result, both distances showed similar results. However, the buffer radius of 400 meters presented mostly better performance.

6.1.5 Calculation of indicators

The flowchart from the Figure 3.2 was followed to calculate the indicators. A sensitivity analysis took place to estimate the SD value that built better models. Since this variable has minor influence in the model performance, four criteria were compared: number of variables, number of collinear variables, average R^2 adjusted for OLS. The logarithmic transformation was also considered. The best threshold value resulted $SD = 5$. Table 6.3 indicates that $SD = 1$ presented the worst results with a lower R^2_{adj} . $SD = 10$ threw similar results as $SD = 5$, but its R^2_{adj} is slightly lower than $SD = 5$. Finally, **194** variables with different indicators were

Table 6.2: Sensitivity analysis to chose the buffer distance

Comparison criteria	Buffer distance [m]	
	400	300
# Variables	200	192
# Not collinear variables	144	137
OLS (Avg. R_{adj}^2)	0.50	0.49
OLS + log (Avg. R_{adj}^2)	0.67	0.66
Stepwise regression +log (Avg. R_{adj}^2)	0.67	0.65
# OLS better models	30	6
# OLS + log better models	35	1
# Stepwise regression +log better models	36	0

+ log: logarithm of the dependent variable

repetitive in the six in the cities of the study.

Table 6.3: Sensitivity analysis to choose the standard deviation threshold

Comparison criteria	SD		
	1	5	10
# Variables	167	194	200
# Not collinear Variables	147	144	144
OLS (Avg. R_{adj}^2)	0.495	0.506	0.501
OLS + log (Avg. R_{adj}^2)	0.666	0.672	0.671

6.1.6 Exploratory Data Analysis (II)

An EDA was performed with the resulting the **36** dependent variables and the **194** independent variables.

– **Dependent variables** – The boxplots from the 36 time units are displayed in Figure and more detailed information is presented in Table 6.4. It is evidenced that there exist a high number of outliers. This is attributed to the fact, that we are using very different cities and also the city centers of the cities attract and produce much more trips than the other zones. On the time intervals in the afternoon, the highest median, mean and standard deviation where observed. On the other hand, the time intervals in the morning on the weekend, and nights on workdays showed the lowest spread, median, and mean, with a mean slightly higher than zero. Finally, Table 6.5 shows a Pearson's correlation between attractions and productions per time unit. The least correlated was the time interval "Afternoon I", then generally arrivals and departures have a high correlation.

– **Independent variables** – The Appendix A shows a summary of the variables including the mean, standard deviation, minimum value, median and the maximum value. To have a panoramic of the main data, the Appendix B shows scatterplots, histograms and Pearson's correlation coefficients between the rentals and the most successful factors named in the literature, the Appendix C between the rentals and the most correlated variables in the dataset. The nomenclature of this variables is based on three parts:

Table 6.4: Summary of the dependent variables

Statistic	Mean	St. Dev.	Pctl(25)	Median	Pctl(75)
WA1p	3.022	4.874	0.210	0.875	3.545
WA2p	4.976	8.123	0.358	1.557	5.716
WM1p	1.955	3.438	0.142	0.528	1.926
WM2p	1.723	2.622	0.159	0.568	2.023
WN1p	1.925	3.260	0.159	0.619	1.943
WN2p	0.443	0.821	0.028	0.125	0.455
SaA1p	3.693	6.665	0.284	1.045	3.369
SaA2p	3.564	6.596	0.267	1.000	3.347
SaM1p	0.384	0.654	0.034	0.119	0.409
SaM2p	1.529	2.629	0.125	0.432	1.483
SaN1p	2.106	3.844	0.165	0.614	2.045
SaN2p	1.743	4.019	0.108	0.415	1.665
SuA1p	3.497	6.387	0.284	0.960	3.205
SuA2p	3.381	6.188	0.278	0.966	3.051
SuM1p	0.339	0.688	0.028	0.091	0.295
SuM2p	1.246	2.175	0.125	0.358	1.153
SuN1p	1.442	2.547	0.125	0.466	1.398
SuN2p	1.849	4.196	0.102	0.449	1.722
WA1a	2.908	4.767	0.188	0.858	3.455
WA2a	5.007	8.381	0.409	1.500	4.892
WM1a	1.732	3.420	0.062	0.386	1.636
WM2a	1.779	3.077	0.102	0.438	2.051
WN1a	2.132	3.488	0.205	0.744	2.153
WN2a	0.486	0.800	0.040	0.159	0.528
SaA1a	3.501	7.029	0.216	0.824	3.136
SaA2a	3.753	6.980	0.267	0.977	3.307
SaM1a	0.353	0.640	0.028	0.097	0.335
SaM2a	1.296	2.354	0.085	0.324	1.222
SaN1a	2.263	4.344	0.170	0.636	2.108
SaN2a	1.854	3.390	0.170	0.619	1.847
SuA1a	3.263	6.506	0.227	0.801	2.847
SuA2a	3.650	6.534	0.312	1.057	3.392
SuM1a	0.324	0.725	0.023	0.080	0.290
SuM2a	1.037	1.853	0.074	0.284	0.966
SuN1a	1.616	2.691	0.153	0.523	1.608
SuN2a	1.863	3.424	0.159	0.574	1.784
Average	2.16	3.89	0.16	0.61	2.12

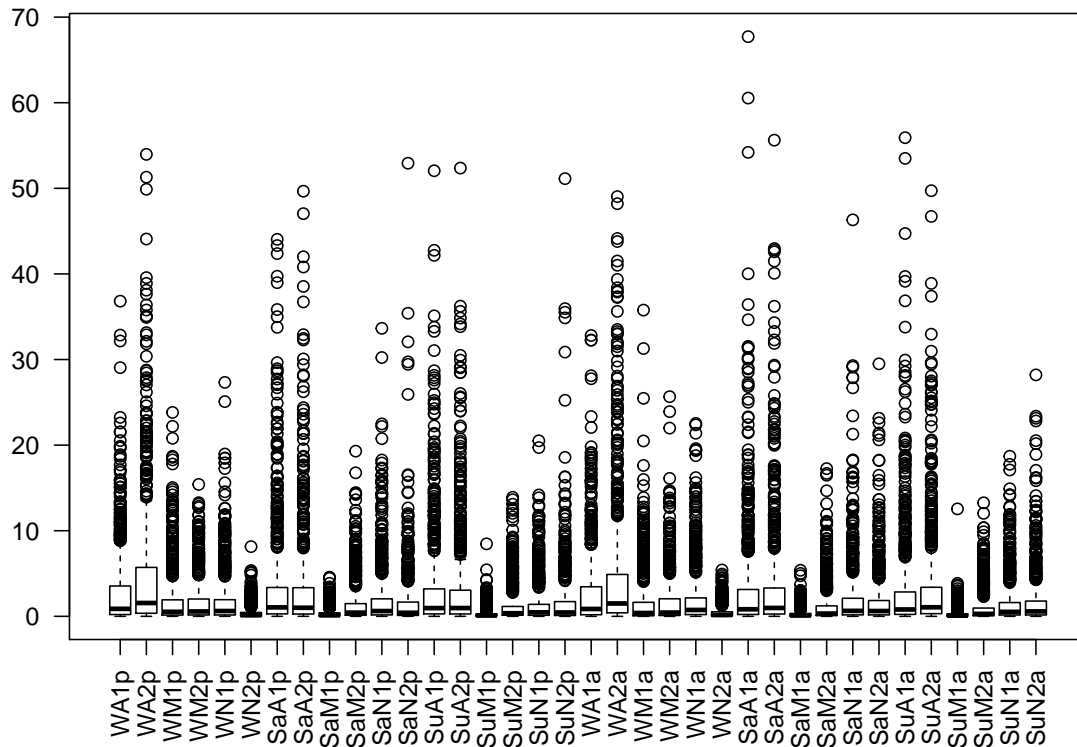


Figure 6.12: Summary of the dependent variables

1. The name of the variable from OSM (<http://wiki.openstreetmap.org>),
2. A code indicating the geometric shape:
 - p=point,
 - l=line,
 - a=area/polygon
3. A code based on the indicator assigned:
 - InArea=prencence in the zone of influence,
 - Distance_min=distance to the feature to the station/centroid in the zone of influence,
 - Distance_min_all=distance from the station/centroid to the closest feature in the city,
 - Density=density of the feature.

For example, *artwork_p_InArea* means the presence in the zone of influence of a public piece of art (as described in OSM) which is presented as a spatial point.

Furthermore, Figure 6.13 shows a summary of the Pearson's correlation coefficients, where we can see the variables that presented the highest coefficients with the rentals, such as the population, clothes stores in the area, memorials in the area and distance to water bodies memorials and densities of trees and cafes. In the same way, Pearson's correlation coefficients were correlated on the variables commonly named in the literature that usually affects the most the deployment of SBBS systems (see Table 2.5. The most correlated variables with the rentals were mainly bars in an area, pubs in an area, cinema in an area, railway stations, restaurants

Table 6.5: Pearson's correlation Attractions vs Productions

Time Unit	Correlation Attractions vs Productions	
1	WA1	0.96
2	WA2	0.9
3	WM1	0.56
4	WM2	0.85
5	WN1	0.91
6	WN2	0.88
7	SaA1	0.96
8	SaA2	0.99
9	SaM1	0.83
10	SaM2	0.86
11	SaN1	0.98
12	SaN2	0.85
13	SuA1	0.98
14	SuA2	0.98
15	SuM1	0.92
16	SuM2	0.9
17	SuN1	0.93
18	SuN2	0.86

among others shown in Figure 6.13 as the population, distance from water bodies, and density of a residential area.

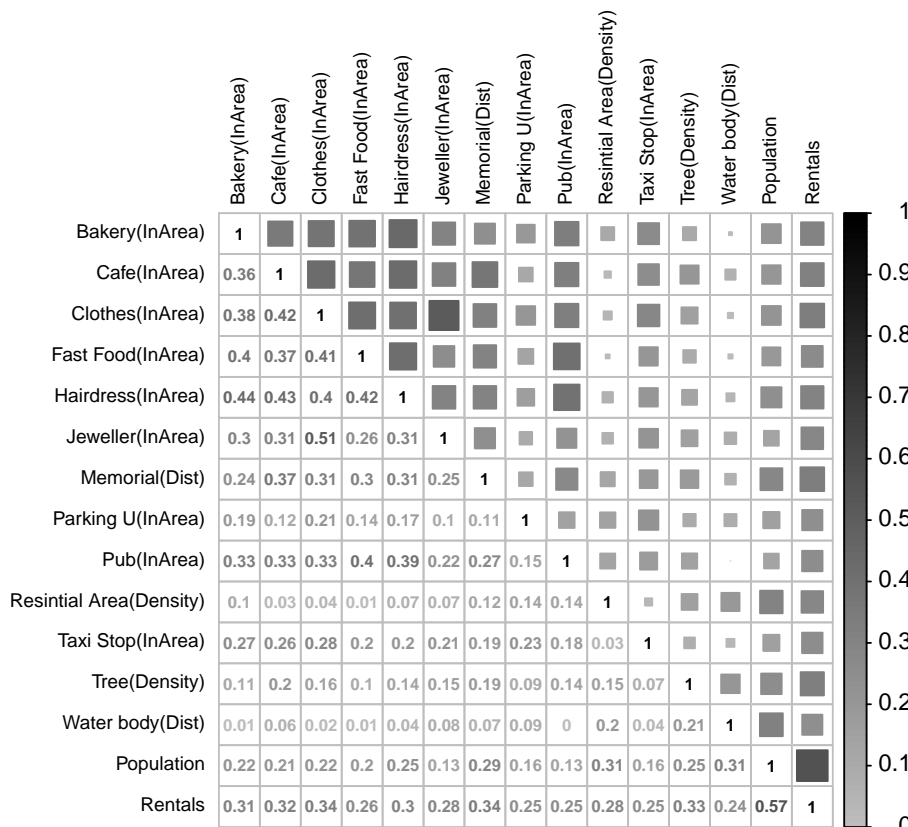


Figure 6.13: Pearson correlation: rentals vs. variables with the highest correlation

Furthermore, a Spearman correlation test was also performed (Figure 6.14) to examine non-linear correlations that might exist. Figure 6.15 shows the correlation coefficients of the rentals and the independent variables using the methods from Pearson vs. Spearman. In other words, points (in this case the independent variables), which are over the 45 degrees line, have a better linear relationship. In the same waypoint under the line have a better monotonic relationship. We can see that much more variables have a monotonic approach (around 70%).

As suggested in the literature (see Section 2.2.5), we can perform some transformations to use Pearson's correlation, in other words, to have a linear approach. Therefore, Table 6.6 shows the results of the average of the correlation coefficients according to different types of transformations. We see that the logarithmic transformation presented the best performance. Then, after a logarithmic transformation of the rentals, we observe that just 35% of the variables behave better with the Spearman's correlation (see Figure 6.16). Thus, after a *logtransformation*, we could use a linear regression method. However, the last plot showed that there was no significant difference in the performance between both correlation methods. Also, the average of the Spearman's correlation is higher. In conclusion, nonlinear regression models might also be implemented.

Table 6.6: Comparison of the performance of transformation of the dependent variable

Transformation	Spearman's coefficient (Average)	Pearson's coefficient (Average)
No Transformation	0.17	0.144
Logarithmic	0.17	0.161
Exponential	0.17	0.046
Square root	0.17	0.159
Square	0.17	0.118
Cube	0.17	0.100
Inverse	0.17	0.008

6.2 Model building and selection

Dependent and independent variables were selected to proceed to the construction of the models. This section presents the results of the built models and also the variables that most influenced them. Collinearity was detected and addressed to construct models following three different regression methods: two linear (stepwise and GLM) and one nonlinear (GBM). Five cities were selected as a training set and a sixth city as a test set. Model diagnostics were carried out to determine transformations of the variables to improve the performance of the models. After these models were validated with the sixth city, they were assessed and compared to determine those that better fitted the dataset. Finally, the variables which were selected by the better-fitted models are aggregated and ranked.

6.2.1 Detecting and addressing collinearity

A collinearity analysis was carried out with the pre-selected variables. The three main criteria to detect collinearity had values much more higher than the acceptable thresholds (see Table 6.7). Therefore, these issues have to be addressed. A sensitivity analysis was realized to determine which technique was better between direct elimination (DE) and VIF elimination. DE had a better performance in criterion I) and II), also it had one variable less than the other technique (see Table 6.7). Furthermore, Table 6.8 shows the R^2 adjusted values of the 36 models using the different 18 times units in both techniques. Also, we calculated the logarithm of the dependent

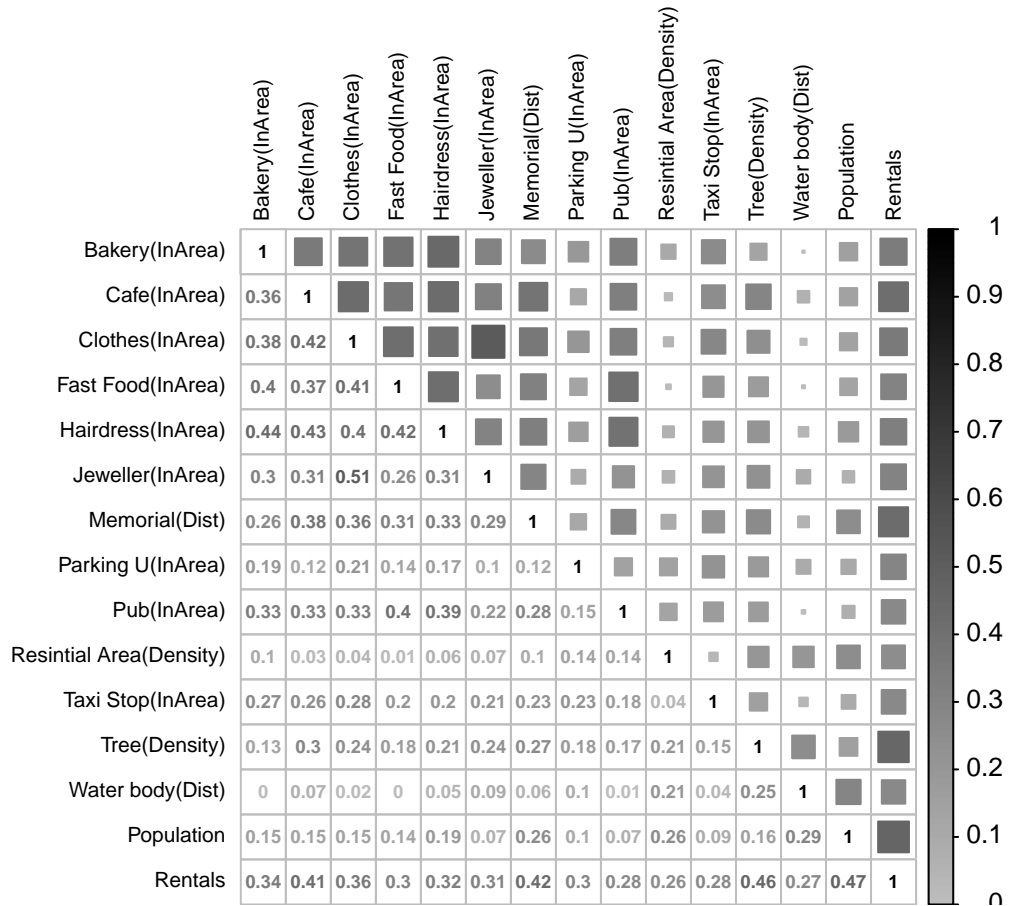


Figure 6.14: Spearman's correlation: rentals vs. variables with the highest correlation

variable as some authors did in the literature (see Table 2.4 and after the results of the EDA. As a result the logarithm dependent variable and the direct elimination technique delivered better results.

Table 6.7: Comparison of the techniques to address collinearity

Technique	# Variables	I) k<100	II) <5	III) #VIFs>5	# Better models	
					Original	Log
Original data	194	1 * 10 ⁶	7 * 10 ⁴	110	-	-
Direct Elimination	144	99.26	1.82	0	33	27
VIF criterion	145	108.29	1.85	0	3	9

6.2.2 Model building and assessment

Since the collinearity between independent variables was removed, the regression models can be estimated. Three regression methods were implemented in the dataset: stepwise regression, GLM, and GBM. They correlated the arrivals and departures aggregated in 18 time units (see Table 6.1) with 144 non-collinear independent variables in the case of stepwise regression, and

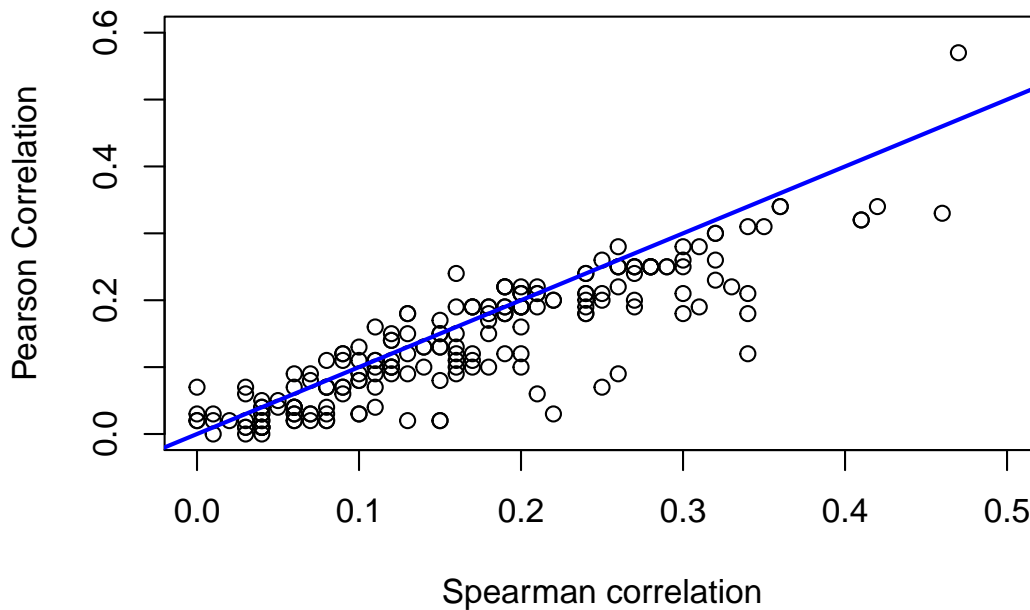


Figure 6.15: Pearson correlation vs spearman correlation of the rentals with the independent variables

194 variables for GLM and GBM.

The ideal case would have been to split the dataset into training, validation and test set, as explained in Section 2.2.1. However, the 689 observations were not a considerable high quantity to break the dataset into these three parts. As a consequence, it was divided into a training set consisting of five cities and validation set using a sixth city. It should be noted that the data were not sampled randomly because the validation purpose was to show how well the models would perform in an additional city with a SBBS system.

Kassel was selected as the city to perform the validation of the models built with the five other cities. Hamburg and Frankfurt could not be removed from the training set because they involved together around the 76% of the zones of influence. Kassel was the next city with more stations after this two cities, and therefore, this city was selected. Another advantage is that a medium-sized city will be validated after a model training with two big-sized, one middle-sized and two small-sized cities.

The parameters R^2 and MSE were considered to measure the performance of the models' validation. R_{adj}^2 and BIC were not used because Kassel has only 58 SBBS stations, which is lower than the number of independent variables used to build the models. This difference is crucial for models that used the GBM regression method because this regression method does not perform a variable selection and therefore, R_{adj}^2 and BIC that have as an input argument the difference between the observations and predictors cannot be estimated.

Moreover, the results of the three regression methods are presented. First, traditional linear models were set with a variable selection technique of stepwise regression in both directions. Then, to analyze if generalized linear models fit better the data, the technique lasso was used to shrinkage the coefficients and perform a variable selection. However, since in the Pearson's and Spearman's correlation analysis some variables showed a monotonic behavior instead of a linear regression (see Section 6.13), a nonlinear regression was implemented by decision trees

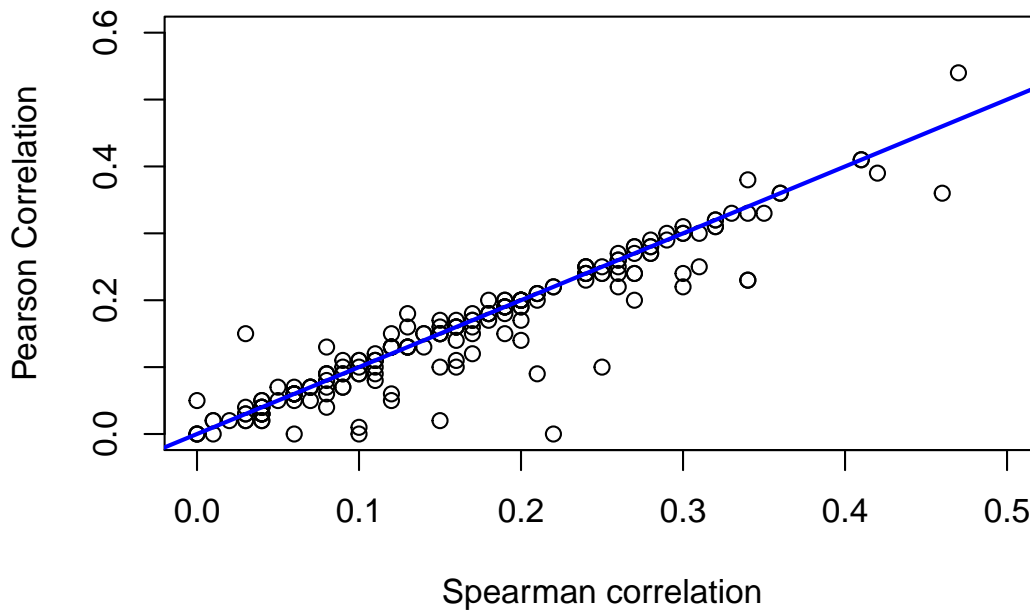


Figure 6.16: Pearson correlation vs spearman correlation after a logarithmic transformation of the rentals with the independent variables

improved with gradient boosting.

The results are presented in four different forms. First, a summary table including the parameters to assess each model (the number of variables selected, indirect values: MSE, R^2 , R^2_{adj} , BIC, validation parameters: MSE and R^2) (see Appendix D). The next two forms are plots showing the fitted vs. the observed values (see Appendix E), and residuals vs. fitted values (see Appendix F) to analyze how well the models fit the data and also to detect outliers, heteroscedasticity and non-normality values and predicted. Finally, plots showing the relationship between predicted and observed data helped to validate the models with the dataset of the city of Kassel (see Appendix G).

– **Stepwise regression** – Linear models with OLS was the first method used to fit the dependent and independent variables. The stepwise regression technique in two directions was carry out to obtain more simple models considering the most influencing variables. It presented an average of 15.47 variables selected in the 36 models and a R^2_{adj} of 0.51 and a R^2 from the validation of 0.20 (see Table 6.9 and Table D.1).

A problems of heteroscedasticity, nonnormality and outliers were present in the models. This statement is based on the findings present in Figure E.1 and Figure F.1 presented in the Appendices. Both figures show that the residuals of the fitted data and observed data grew when the observed data increased (see an example in Figure 6.18). Also, we can see in Figure E.1 that the data did not fit a tendency of a 45-degree line, but instead they presented a concave shape. Therefore, a transformation of the dependent variable by using a concave function as the logarithm or the square root could solve the heteroscedasticity. A logarithmic (log) transformation was implemented because it showed a higher correlation between the dependent and independent variable (see Table 6.6). A boxcox transformation was also applied because of the recommendation in the literature of its high performance (Box and Cox, 1964).

Table 6.8: Sensitivity analysis to choose the best way to address collinearity

Model	DE (R_{adj}^2)	VIF (R_{adj}^2)	DE+log (R_{adj}^2)	VIF+log (R_{adj}^2)	DE > VIF ($TRUE=1$)	DE+log>VIF+log ($TRUE=1$)
WA1p	0.56	0.56	0.69	0.69	1	0
WA2p	0.59	0.59	0.70	0.70	1	0
WM1p	0.47	0.47	0.63	0.63	1	1
WM2p	0.56	0.56	0.68	0.68	1	1
WN1p	0.53	0.52	0.66	0.66	1	0
WN2p	0.54	0.54	0.66	0.66	1	0
SaA1p	0.51	0.51	0.72	0.72	1	1
SaA2p	0.50	0.50	0.71	0.71	1	0
SaM1p	0.49	0.49	0.59	0.59	1	1
SaM2p	0.50	0.50	0.66	0.66	1	1
SaN1p	0.53	0.53	0.69	0.70	1	0
SaN2p	0.46	0.46	0.68	0.68	1	1
SuA1p	0.49	0.49	0.71	0.71	1	1
SuA2p	0.47	0.47	0.70	0.70	1	1
SuM1p	0.44	0.43	0.62	0.62	1	1
SuM2p	0.49	0.49	0.63	0.63	1	1
SuN1p	0.52	0.52	0.68	0.68	1	1
SuN2p	0.47	0.47	0.72	0.72	1	1
WA1a	0.55	0.55	0.69	0.69	1	0
WA2a	0.53	0.53	0.68	0.68	1	1
WM1a	0.45	0.45	0.63	0.63	1	1
WM2a	0.55	0.55	0.68	0.68	1	1
WN1a	0.53	0.53	0.63	0.63	0	1
WN2a	0.53	0.53	0.61	0.61	1	1
SaA1a	0.48	0.47	0.72	0.72	1	1
SaA2a	0.51	0.51	0.71	0.71	1	1
SaM1a	0.44	0.43	0.62	0.62	1	0
SaM2a	0.50	0.50	0.70	0.70	1	1
SaN1a	0.53	0.53	0.69	0.69	1	1
SaN2a	0.50	0.50	0.65	0.65	0	1
SuA1a	0.44	0.44	0.71	0.71	1	1
SuA2a	0.50	0.50	0.70	0.70	1	1
SuM1a	0.33	0.33	0.61	0.60	1	1
SuM2a	0.48	0.48	0.67	0.67	1	1
SuN1a	0.54	0.54	0.66	0.66	1	1
SuN2a	0.51	0.51	0.66	0.66	0	0
Average	0.500	0.499	0.670	0.670	0.92	0.75

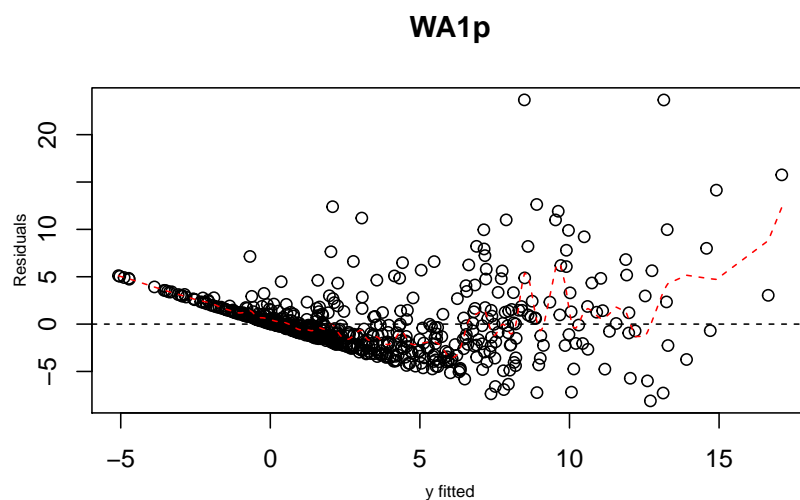
DE=Direct Elimination, VIF= Variance Inflation Factor
 log= logarithm of the dependent variable

Commonly, outliers were present in zones of influence with zero arrivals or departures in a time interval. Thus, this zones with zero values were removed from the dependent variables to reduce the outliers problem and also to have real numbers after their logarithmic and boxcox transformations.

As a result of the transformations, Table 6.9 indicates that as an average, a considerable increase of around 0.2 in the R_{adj}^2 was present after the logarithmic and boxcox transformation and also, a decrease of the MSE of around that ten times. Figure E.2 and Figure E.3, and Figure F.2 and Figure F.3 presented in the appendix show that respectively log and boxcox presented models with fitted values more similar than the observed and a relatively more equal distribution of the values, making the model's results more homoscedastic (see an example in Figure). Generally, log and boxcox transformations presented similar results (see Table 6.9).

Table 6.9: Comparison of the results from the Stepwise regression

Assessment method	No transformation			Logarithmic transformation			BoxCox transformation		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
N. of variables	15.47	8	21	16.56	13	22	17.64	13	25
MSE	10.17	0.24	33.13	0.98	0.85	1.44	0.40	0.009	1.724
R^2	0.52	0.37	0.59	0.70	0.63	0.75	0.72	0.643	0.777
R^2_{adj}	0.51	0.36	0.59	0.69	0.62	0.75	0.71	0.632	0.77
BIC	2860	929	4096	1843	1709	2048	1149	283	2250
MSE (validation)	5.64	0.12	17.18	2.54	1.54	3.68	1.32	0.02	4.38
R^2 (validation)	0.20	0.07	0.44	0.41	0.14	0.65	0.40	0.14	0.66

**Figure 6.17:** Example of heteroscedacity (Stepwise regression)

However, in average parameters from the indirect methods showed a better performance of the boxcox transformation. After these close relationship between the two transformations, a decision to choose the one that better performed in a general form was not possible. A further analysis was required by comparing the R from the validation in each time interval to choose the transformation that displayed the better-fitted data. The R^2 of the validation was chosen because it is a more direct way to understand how model would behave in another city. The average difference of the R^2 from the validation between both transformation was 0.025. but 22 of the 36 models presented a better validation R^2 by calculating the logarithm of the dependent variable (see Table D.2 and Table D.3).

According to the variables selection, the original dataset presented as average one variable less than the log transformation and two less than the boxcox transformation. As an example Table 6.10, Table 6.11 and Table 6.12 show a summary of six different results of each type of transformation and their variables selected. Even though log and boxcox transformed models selected a relatively higher number of variables than the original dataset, they presented a more logical selection of the variables. For instance, cinemas and nightclubs were more representative on Saturday night, cafés on Workday morning and water areas on Sunday afternoons.

In conclusion, boxcox and log transformations presented a better variable selection and also a better fitted and validated models than the original dataset.

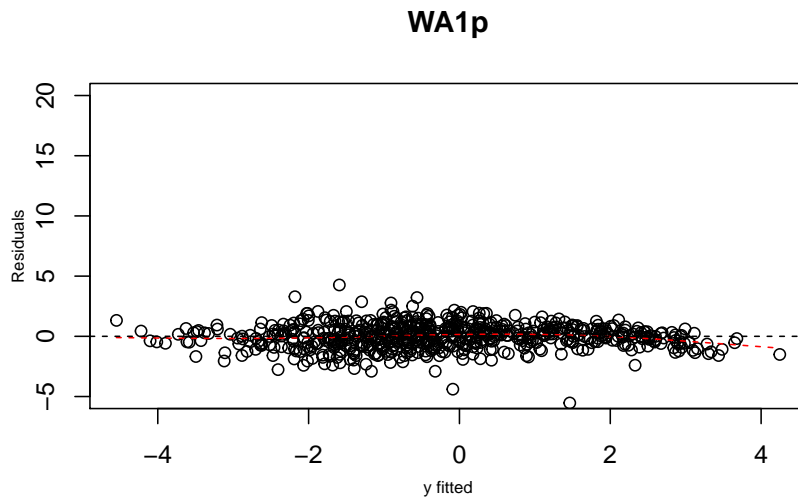


Figure 6.18: Example of homoscedacity (Stepwise regression with logarithmic transformation)

– **Generalized linear models + lasso** – The second approach was the generalized linear models using lasso a shrinkage technique. The hypothesis was that the data might behave better if the errors are fitted to a distribution function than with OLS. To build the GLM models, a Gaussian distribution was considered because this family according to the literature is useful for normal distributed errors (see Section 2.2.3) as the dataset is expected to have. Moreover, other distribution did not fit on the trained data. As λ is the main input argument to run these models, a k-folds cross validation was implemented to calculate the λ that help the models to fit better the data. The number of folds were set equal to the number of independent values (in this case 194), which is the highest possible number of iterations to have the best performance of the models.

The output presented the same issues as in the stepwise regression results of outliers and heteroscedasticity (see Figure E.4 and Figure F.4). Therefore, the dependent variables were also transformed with logarithmic and boxcox functions solving the heteroscedasticity issue (see Figure E.5, Figure F.5 (log); & Figure E.6, Figure F.6 (boxcox)). An additional issue detected was the outcome in the time interval "SuMa1", after its low demand GLM shrank the model to zero variables. But this issue was also solved after the respective transformations.

An improvement of an average around 0.3 in the R_{adj}^2 and 0.22 R^2 of the validation were present in the models after the transformations. As in stepwise regression the results between log and boxcox were very similar. However, 23 out of 36 models fitted better the dataset after a logarithmic transformation. However, the transformed models selected around the double more variables than the original dataset (around 15 as average). But R^2 from the validation was much higher the variables from the transformed models are preferred. A summary of the results is presented in Table 6.13.

Table 6.10: Stepwise regression results (No transformation)

	WA2a	WM1a	WN1a	SaN1a	SuA2a	SuM2a
bakery_p_Distance_min	-0.00002*** (0.00001)	-0.00001*** (0.00000)	-0.00001*** (0.00000)			
cinema_p_InArea	3.879*** (1.073)		2.674*** (0.462)	3.680*** (0.591)	2.377*** (0.884)	0.917*** (0.253)
clothes_p_Distance_min					-0.00002*** (0.00001)	
cycleway_l_Distance_min			-0.00001*** (0.00000)		-0.00001*** (0.00000)	
footway_l_Density			-0.00004*** (0.00002)			
computer_shop_p_Distance_min	-0.00003*** (0.00001)		-0.00001** (0.00000)	-0.00002*** (0.00000)		
fountain_p_Distance_min	0.00002*** (0.00001)					
florist_p_Distance_min				-0.00001*** (0.00000)		
guesthouse_p_InArea				-2.265*** (0.839)		
department_store_p_InArea					-3.570*** (1.039)	
jeweller_p_InArea					1.707*** (0.634)	0.490*** (0.167)
library_p_InArea					-1.562*** (0.524)	
nightclub_p_InArea	2.598** (1.008)		1.618*** (0.424)	3.567*** (0.542)	2.390*** (0.814)	
outdoor_shop_p_InArea			-2.169*** (0.625)	-3.374*** (0.817)		
parking_multistorey_a_Density	-80.707** (31.829)		-42.224*** (13.094)	-56.386*** (16.587)	-93.490*** (25.808)	-26.401*** (7.514)
bank_p_Distance_min		-0.00001*** (0.00000)				
car_rental_p_InArea		1.262*** (0.428)				
chemist_p_InArea		-0.959*** (0.316)				
path_l_Distance_min				-0.00001*** (0.00000)		
pedestrian_l_Density	-0.0002*** (0.00005)	0.0001*** (0.00002)	-0.0001*** (0.00002)	-0.0001*** (0.00002)	-0.0001*** (0.00004)	
pharmacy_p_Distance_min	-0.00002*** (0.00001)				-0.00001*** (0.00000)	
picnic_site_p_InArea		3.631*** (0.827)				
pitch_p_InArea				1.802*** (0.551)		
pub_p_InArea			0.716*** (0.231)	0.872*** (0.288)		
theatre_p_Distance_min				-0.00001*** (0.00000)		
residential_a_Density	-2.794*** (0.761)	-2.059*** (0.356)	-1.204*** (0.412)	-1.800*** (0.522)		
steps_l_Density	0.001*** (0.0002)					
taxi_p_Distance_min	-0.00002*** (0.00001)					-0.00000*** (0.00000)
residential_l_Density			0.0001*** (0.00003)	0.0001*** (0.00004)		
tertiary_l_Distance_min			-0.00001*** (0.00000)	-0.00001*** (0.00000)		
tree_p_Density	0.004*** (0.001)		0.002*** (0.0004)		0.003*** (0.001)	
university_p_InArea	3.703*** (1.164)	1.445*** (0.533)	1.520*** (0.485)	1.763*** (0.620)	2.743*** (0.944)	0.946*** (0.275)
water_a_Distance_min					-0.00001*** (0.00000)	
City_center_Distance_min_all	-0.001*** (0.0001)	-0.0004*** (0.00004)	-0.0004*** (0.00004)	-0.0004*** (0.00005)	-0.001*** (0.0001)	-0.0002*** (0.00002)
Population	0.00001*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00001*** (0.00000)	0.00000*** (0.00000)
scrub_a_Distance_min						-0.00000*** (0.00000)
Constant	5.825*** (1.370)	2.174*** (0.465)	1.543*** (0.561)	3.654*** (0.834)	2.189** (0.858)	0.279 (0.203)
Observations	629	598	626	622	629	610
R ²	0.560	0.450	0.558	0.551	0.521	0.483
Adjusted R ²	0.549	0.441	0.546	0.537	0.509	0.476
Residual Std. Error	5.830 (df = 613)	2.701 (df = 587)	2.425 (df = 608)	3.077 (df = 603)	4.744 (df = 613)	1.404 (df = 601)
F Statistic	51.917*** (df = 15; 613)	48.040*** (df = 10; 587)	45.133*** (df = 17; 608)	41.091*** (df = 18; 603)	44.485*** (df = 15; 613)	70.153*** (df = 8; 601)

Note:

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

Table 6.11: Stepwise regression (Logarithmic transformation)

	WA2a	WM1a	WN1a	SaN1a	SuA2a	SuM2a
allotments_a.Distance_min		0.0000*** (0.00000)				
artwork_p.InArea	0.434*** (0.101)	0.410*** (0.119)	0.464*** (0.104)	0.289*** (0.097)	0.427*** (0.097)	0.353*** (0.095)
bakery_p.Distance_min	-0.00001*** (0.00000)	-0.00000*** (0.00000)	-0.00001*** (0.00000)	-0.00001*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)
car_sharing_p.Distance_min	-0.00000*** (0.00000)	-0.00000*** (0.00000)				
car_rental_p.InArea			0.401** (0.156)	0.441*** (0.150)	0.404*** (0.148)	0.468*** (0.151)
commercial_a.Distance_min	-0.00000** (0.00000)		-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	
crossing_p.Distance_min					0.00000*** (0.00000)	
cycleway_l.Density	0.00005*** (0.00002)		0.0001*** (0.00002)		0.0001*** (0.00002)	
forest_a.Distance_min					-0.00000*** (0.00000)	
fountain_p.Distance_min	0.00000*** (0.00000)		0.00000*** (0.00000)		0.00000*** (0.00000)	
fuel_a.Distance_min				-0.00000** (0.00000)		
hairdresser_p.Distance_min	-0.00000*** (0.00000)		-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)
farm_a.Density		-4.286*** (1.538)				
kindergarten_p.InArea		-0.349*** (0.105)				
memorial_p.Distance_min			-0.00000*** (0.00000)	-0.00000** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)
nightclub_p.InArea				0.428*** (0.152)		
park_a.Density				0.952** (0.373)	1.276*** (0.385)	
parking_bicycle_a.Density						161.843*** (51.485)
parking_bicycle_p.Distance_min	-0.00000*** (0.00000)	-0.00001*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)
pedestrian_l.Density	-0.00002*** (0.00001)		-0.00003*** (0.00001)			
pub_p.InArea	0.258*** (0.092)					
rail_l.Density		0.00002*** (0.00001)				
residential_a.Density	-0.709*** (0.162)	-0.761*** (0.186)				
residential_a.Distance_min	-0.00000*** (0.00000)		-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00001*** (0.00000)
residential_l.Density	0.00004*** (0.00001)		0.0001*** (0.00001)	0.00004*** (0.00001)	0.00004*** (0.00001)	
scrub_a.Density	-4.675*** (1.447)					
steps_l.Density	0.0001*** (0.00004)					
shelter_p.InArea			0.471*** (0.135)	0.592*** (0.129)	0.446*** (0.128)	0.474*** (0.133)
supermarket_p.Distance_min					-0.00000*** (0.00000)	
traffic_signals_p.Distance_min						0.00000*** (0.00000)
tree_p.Density	0.001*** (0.0002)		0.001*** (0.0002)	0.0005*** (0.0001)	0.0004*** (0.0001)	0.0005*** (0.0002)
turning_circle_p.Distance_min			-0.00000** (0.00000)		-0.00000*** (0.00000)	
university_p.InArea	0.584*** (0.196)		0.606*** (0.200)	0.555*** (0.191)		0.616*** (0.194)
water_a.Density					4.592*** (1.536)	4.592*** (1.496)
City_center.Distance_min_all	-0.0003*** (0.00002)	-0.0003*** (0.00002)	-0.0003*** (0.00002)	-0.0003*** (0.00002)	-0.0003*** (0.00002)	-0.0003*** (0.00002)
Population	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)
service_l.Density		0.0001*** (0.00002)				
commercial_a.Density		1.505*** (0.423)				
Constant	0.306 (0.223)	-1.328*** (0.260)	-0.595*** (0.222)	-0.321 (0.216)	-0.131 (0.213)	-1.449*** (0.129)
Observations	687	653	684	680	687	668
R ²	0.699	0.634	0.647	0.701	0.720	0.680
Adjusted R ²	0.691	0.626	0.637	0.693	0.711	0.672
Residual Std. Error	1.004 (df = 667)	1.243 (df = 639)	1.032 (df = 665)	0.984 (df = 662)	0.971 (df = 665)	1.007 (df = 652)
F Statistic	81.589*** (df = 19; 667)	85.056*** (df = 13; 639)	67.711*** (df = 18; 665)	91.236*** (df = 17; 662)	81.457*** (df = 21; 665)	92.183*** (df = 15; 652)

Note:

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

Table 6.12: Stepwise regression results (Boxcox transformation)

	WA2a	WM1a	WN1a	SaN1a	SuA2a	SuM2a
allotments_a.Distance_min		0.0000*** (0.00000)				
arts_centre_p.InArea		-0.747*** (0.232)				
artwork_p.InArea	0.347*** (0.107)	0.394*** (0.127)	0.363*** (0.110)		0.314*** (0.102)	0.281*** (0.099)
bakery_p.Distance_min	-0.00000*** (0.00000)	-0.00001*** (0.00000)	-0.00001*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00001*** (0.00000)
car_rental_p.InArea				0.385** (0.152)	0.395*** (0.149)	0.387** (0.153)
car_sharing_p.Distance_min	-0.00001*** (0.00000)	-0.00001*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)
cinema_p.InArea				0.466*** (0.177)		
commercial_a.Distance_min	-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	
crossing_p.Distance_min					0.00000*** (0.00000)	
cycleway_l.Density	0.0001*** (0.00002)		0.0001*** (0.00002)	0.0001*** (0.00002)	0.0001*** (0.00002)	0.00005*** (0.00002)
fountain_p.Distance_min	0.00000*** (0.00000)		0.00000*** (0.00000)		0.00000*** (0.00000)	
farm_a.Density		-4.465*** (1.526)				
footway_l.Density		-0.00003*** (0.00001)				
forest_a.Distance_min		0.00000*** (0.00000)				
kindergarten_p.InArea		-0.297*** (0.108)				
mobile_phone_shop_p.Distance_min	-0.00000*** (0.00000)	-0.00000*** (0.00000)				
optician_p.Distance_min		0.00001*** (0.00000)				
hairdresser_p.Distance_min				-0.00000*** (0.00000)	-0.00000*** (0.00000)	
memorial_p.Distance_min			-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)
park_a.Density				1.133*** (0.382)	1.392*** (0.394)	
parking_bicycle_a.Density						154.960*** (50.691)
parking_bicycle_p.Distance_min	-0.00000*** (0.00000)	-0.00001*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00000*** (0.00000)
pedestrian_l.Density	-0.00002*** (0.00001)		-0.00002*** (0.00001)			
pub_p.InArea	0.323*** (0.092)		0.398*** (0.094)	0.306*** (0.092)		
rail_l.Density		0.00002*** (0.00001)				
residential_a.Density	-0.567*** (0.164)	-0.806*** (0.209)				
residential_a.Distance_min					-0.00000** (0.00000)	-0.00000*** (0.00000)
residential_l.Density	0.00005*** (0.00001)	-0.00004*** (0.00002)	0.00005*** (0.00001)	0.00004*** (0.00001)	0.00005*** (0.00001)	
scrub_a.Density	-4.900*** (1.443)					
steps_l.Density	0.0001*** (0.00004)					
shelter_p.InArea				0.430*** (0.158)	0.395*** (0.151)	
supermarket_p.Distance_min					-0.00000** (0.00000)	
turning_circle_p.Distance_min					-0.00000*** (0.00000)	
traffic_signals_p.Distance_min						0.00000*** (0.00000)
university_p.InArea	0.673*** (0.199)	0.677*** (0.247)	0.706*** (0.203)	0.741*** (0.191)	0.622*** (0.188)	0.740*** (0.194)
water_a.Density					4.272*** (1.501)	
City_center.Distance_min_all	-0.0003*** (0.00002)	-0.0003*** (0.00002)	-0.0003*** (0.00002)	-0.0003*** (0.00002)	-0.0003*** (0.00001)	-0.0003*** (0.00002)
Population	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)
cafe_p.InArea	0.241** (0.095)					
Constant	0.008 (0.258)	-0.617** (0.311)	-1.032*** (0.236)	-0.859*** (0.215)	-0.469** (0.229)	-1.581*** (0.153)
Observations	629	598	626	622	629	610
R ²	0.722	0.663	0.660	0.726	0.751	0.709
Adjusted R ²	0.714	0.652	0.652	0.718	0.742	0.702
Residual Std. Error	0.983 (df = 610)	1.219 (df = 578)	1.020 (df = 611)	0.962 (df = 605)	0.940 (df = 607)	0.979 (df = 596)
F Statistic	88.153*** (df = 18; 610)	59.923*** (df = 19; 578)	84.690*** (df = 14; 611)	99.961*** (df = 16; 605)	87.016*** (df = 21; 607)	111.454*** (df = 13; 596)

Note:

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

Table 6.13: Comparison of the results from the GLM regression

Assessment method	No transformation			Logarithmic transformation			BoxCox transformation		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
N. of variables	15	0	35	30	21	42	33	24	50
MSE	13.40	0.29	39.87	1.09	0.94	1.57	0.47	0.01	1.94
R^2	0.36	0.00	0.51	0.67	0.59	0.73	0.69	0.61	0.74
R^2_{adj}	0.34	0.00	0.48	0.65	0.57	0.71	0.67	0.59	0.72
BIC	1424	293	2467	243	125	492	701	10	2286
MSE (validation)	2.67	0.05	9.25	2.11	1.30	3.07	1.10	0.02	3.06
R^2 (validation)	0.22	0.03	0.43	0.44	0.15	0.69	0.44	0.15	0.70

– **GBM** – The third and last approach was the gradient boosting machine. This machine learning technique was implemented in the dataset to analyze its possible nonlinear behavior. As mentioned on the methodological framework the input attributes were set as conservative as possible. A k-fold cross-validation was realized to fine the better number of trees or interations with an input of 5 folds, a shrinkage factor of 0.0001, and a interaction depth of 6. The results of this method did not present a significant heteroscedasticity (see Figure E.7 and Figure F.7. However, logarithmic and boxcox transformations were also realized to analyzed the case if the models fitted better the dataset with them. The logarithmic transformation performed better in 21 of the 36 models. As in the two previous methods, the R^2 from the validation presented a significant increase of around 0.2. However, unlike the linear methods the average of the adjusted R^2 remained constant. A summary of this results are shown in Table 6.14.

Table 6.14: Comparison of the results from the GBM regression

Assessment method	No transformation			Logarithmic transformation			BoxCox transformation		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
N. of trees	7523.14	4545	9529	6127.00	4379	9836	5945	3976	8014
MSE	1.29	0.24	3.25	0.63	0.48	0.74	0.36	0.07	0.85
R^2	0.88	0.59	0.94	0.88	0.82	0.93	0.88	0.82	0.93
R^2_{adj}	0.84	0.45	0.93	0.84	0.76	0.91	0.85	0.76	0.9
BIC	1161	105	2402	346	3	555	701	21	2171
MSE (validation)	2.54	0.05	9.31	1.13	0.78	1.62	0.57	0.01	1.58
R^2 (validation)	0.23	0.04	0.47	0.47	0.21	0.72	0.46	0.2	0.72

Regarding the variables selection, the outcome of this regression method is a ranking list order by the relative influence of the variables on the model instead of a variables selection. Figure 6.19 shows an example of the resulting ranking list of GBM. From this list, an analysis was carried out to determine the most significant variables. MSEs were calculated from sets of variables starting from a set with the highest ranked and then adding a subsequent variable until a non-significant difference of the MSE was present. Each time interval behaves different so the difference threshold of the MSEs changes as well. Figure 6.20, Figure 6.21 and Figure 6.22 show the thresholds for each time interval and for each transformation.

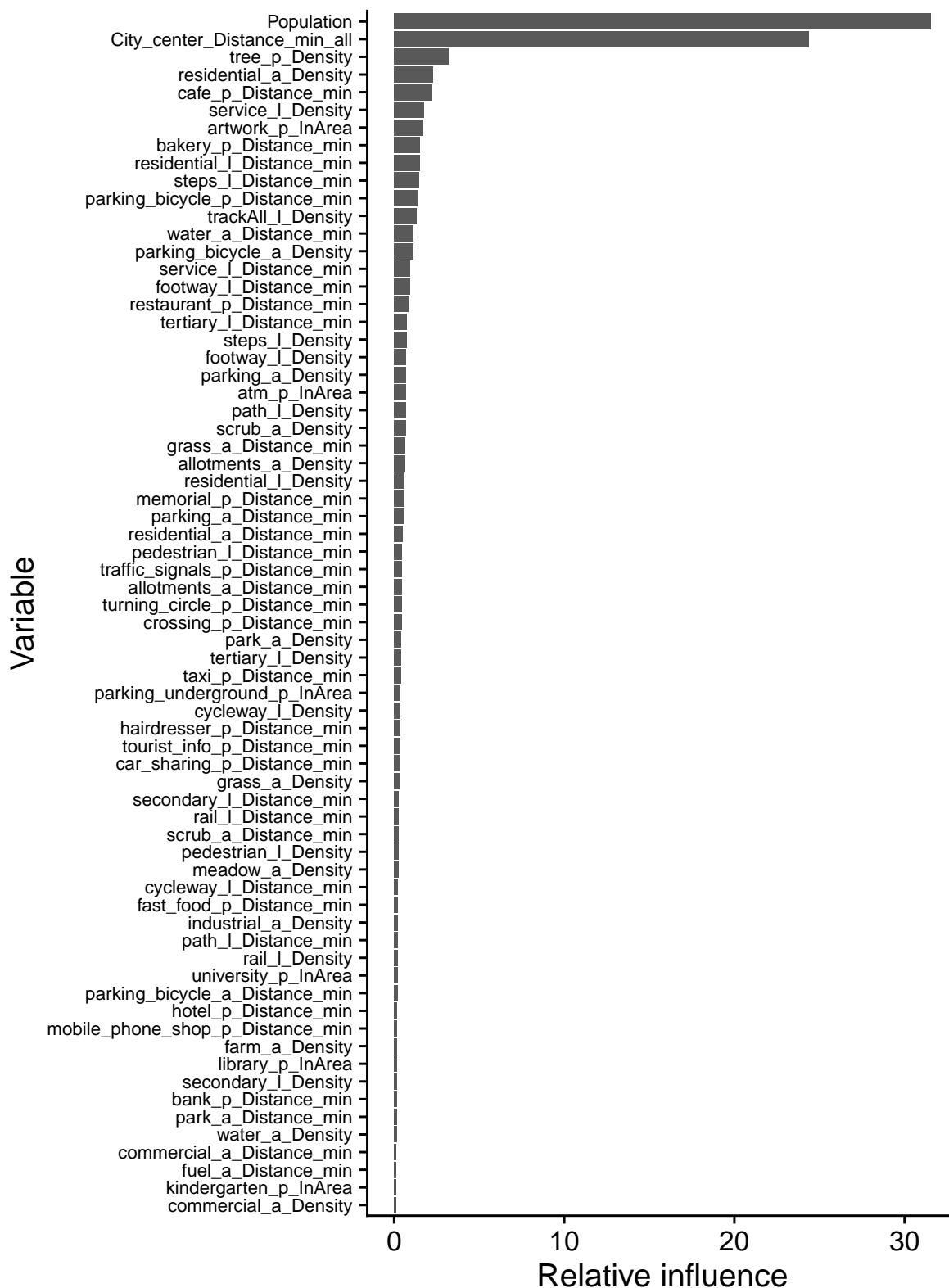


Figure 6.19: Example of GBM relative influence of variables ("WA1p" with BoxCox transformation)

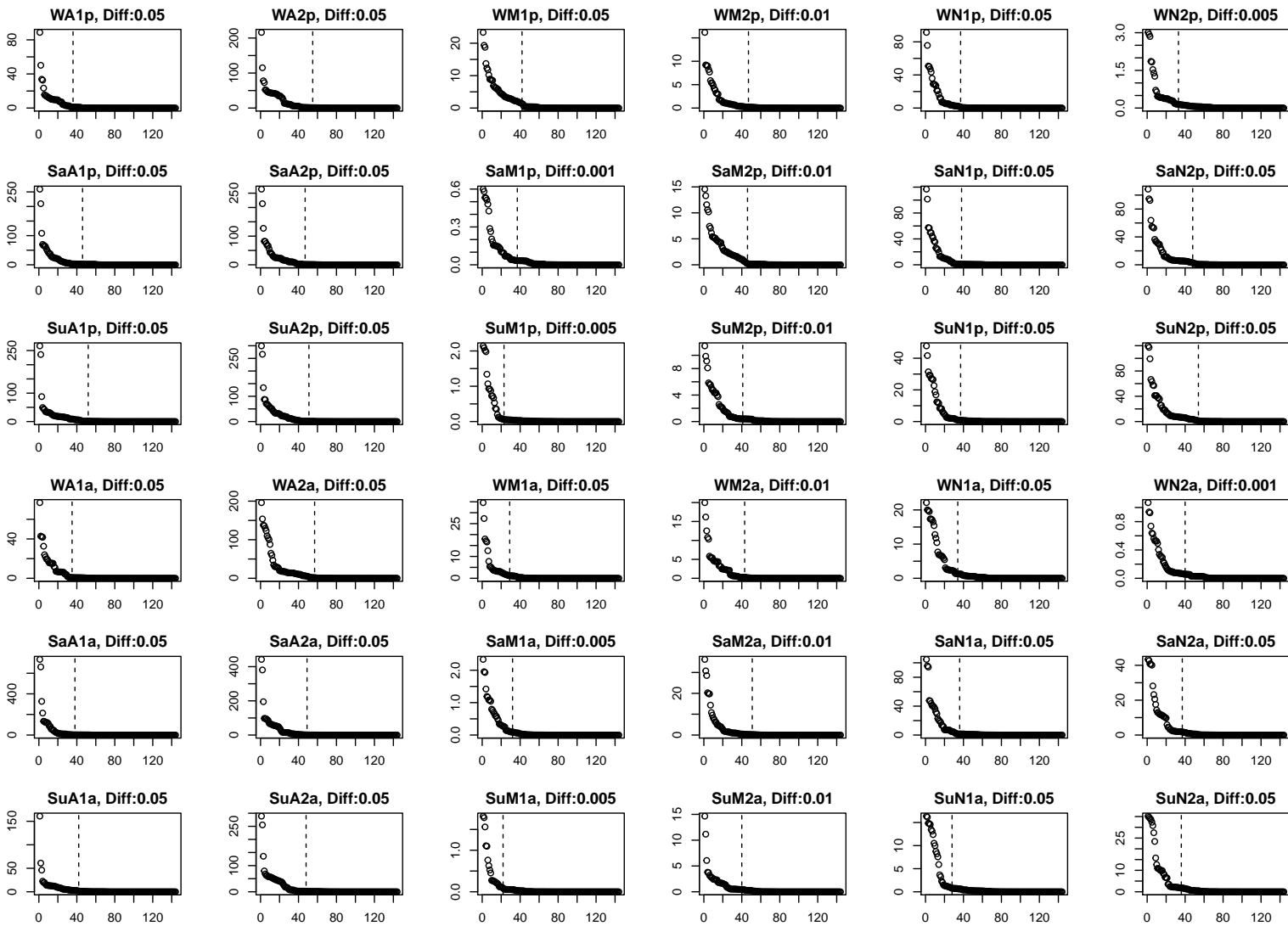


Figure 6.20: Sensitivity analysis to set a threshold for the variables selection in GBM (No transformations)

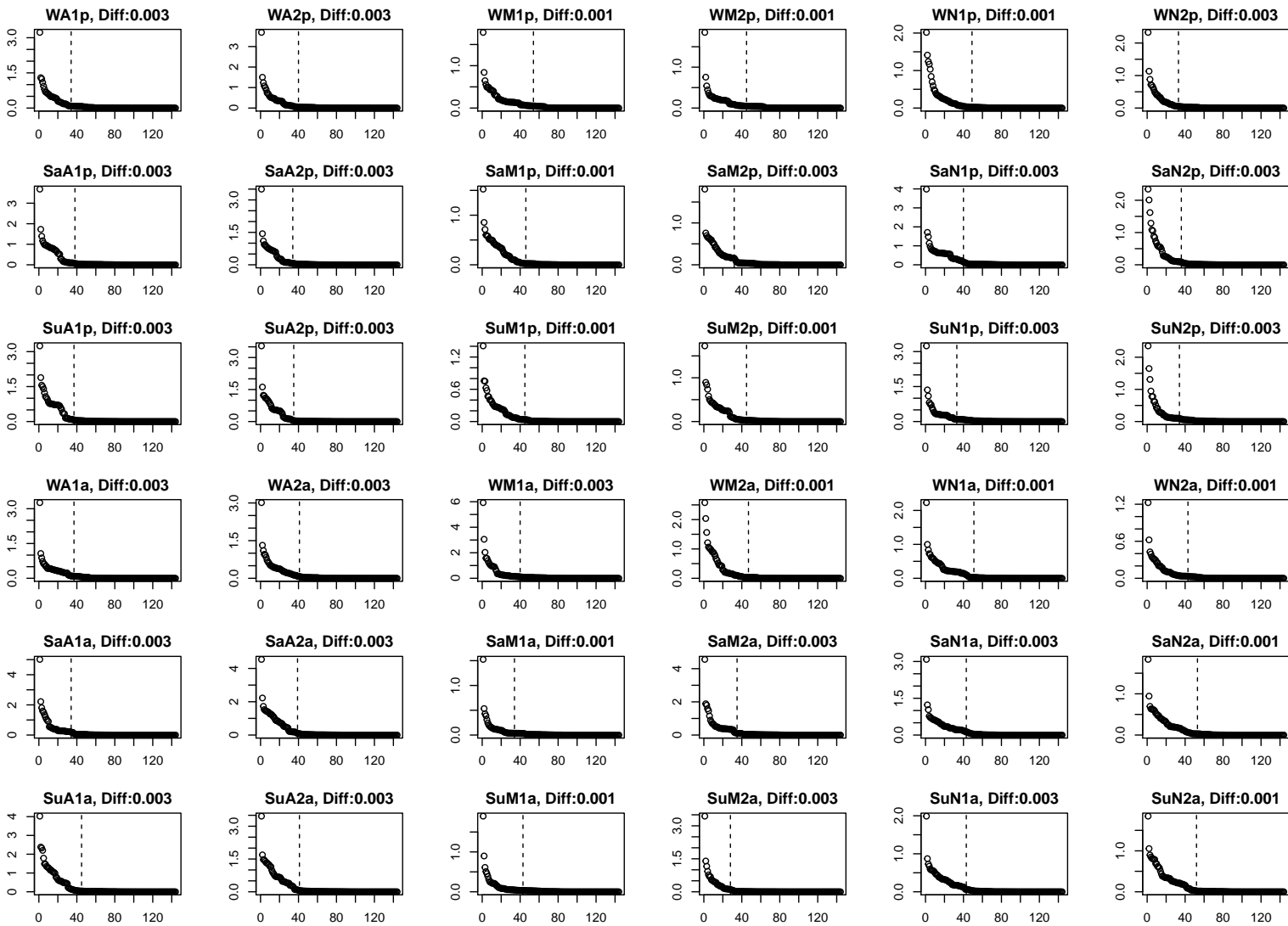


Figure 6.21: Sensitivity analysis to set a threshold for the variables selection in GBM (Log transformation)

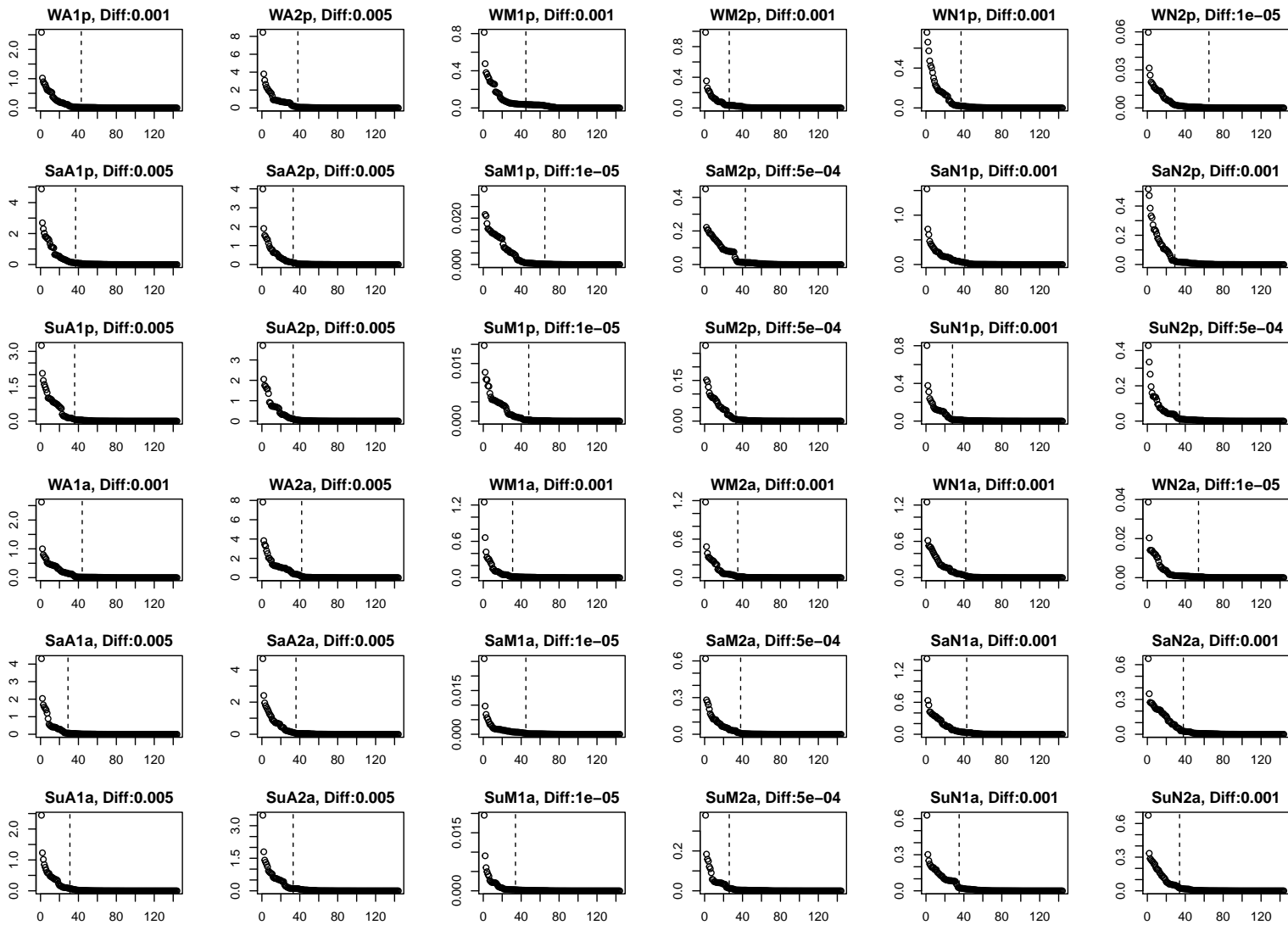


Figure 6.22: Sensitivity analysis to set a threshold for the variables selection in GBM (Boxcox transformation)

6.2.3 Models assessment and selection

In total 324 models were built. 36 models including arrivals and departures per time interval per regression method and transformation technique. The task in this section was to analyze those models and establish those that better fitted the dataset. The two criteria were considered to select the the models are going to be evaluated: 1) the model that better-fitted the data and 2) the model that is more parsimonious.

The analyze the first criteria, the data fitting the models was assessed by the R^2 from the validation and the R^2_{adj} for the models' fitting. R^2_{adj} results were relatively similar in all the times intervals per each regression method. There is not much difference between the arrivals and departures models. The three regression methods presented a relative "parallelism" between the values. This statement means that the models that better fitted the dataset in the stepwise regression also did it in GLM and GBM and also, those that fitted the worst of them (see Figure 6.24). GBM presented the highest values followed by GLM. Between GLM and stepwise regression, there was not a significant difference as comparing GBM and GLM. Almost all time intervals presented a uniform values per regression method. But a significant low performance was in the morning periods.

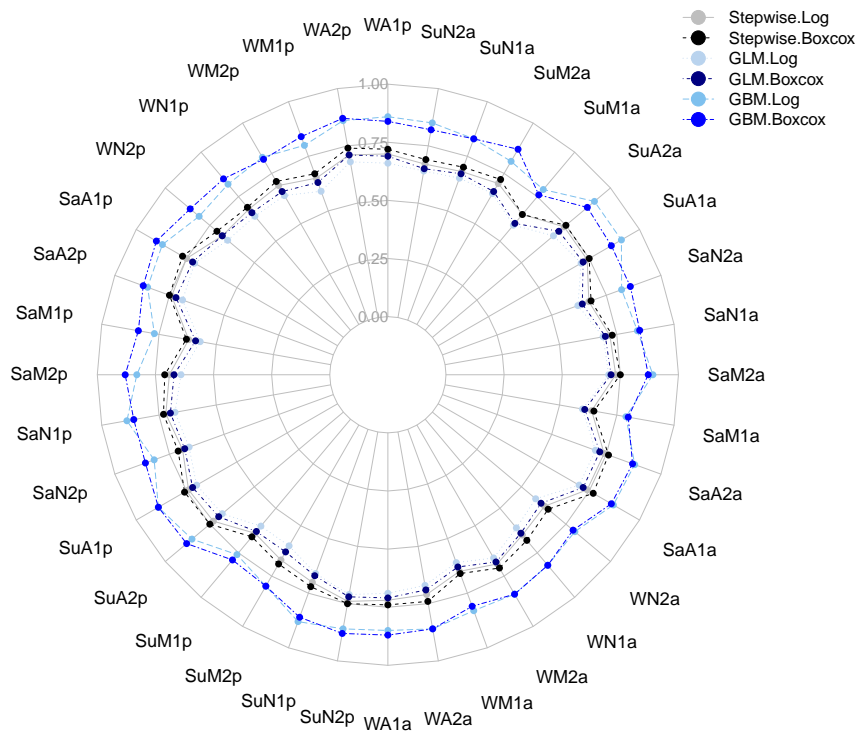


Figure 6.23: Comparison of the R^2 adjusted from the fitted values of different models

On the other hand, R^2 from the validation did not present a significant gap between models and relative same fitting between arrivals and departures models, but the relative "parallelism" remains (see Figure 6.24). As for R^2_{adj} , GBM presented usually the highest validation performance. Arrivals and departures models behaved different especially on Sundays. The better validated R^2 was at the night, while the worst was in the morning.

According to the transformation of the dependent variables, in the previous section they were needed for a better fitting of the models. The three regression methods showed higher $R^2_{validated}$ after a logarithmic or boxcox transformation. However, between these both transformations, a selection was not possible because their better performance differed according to

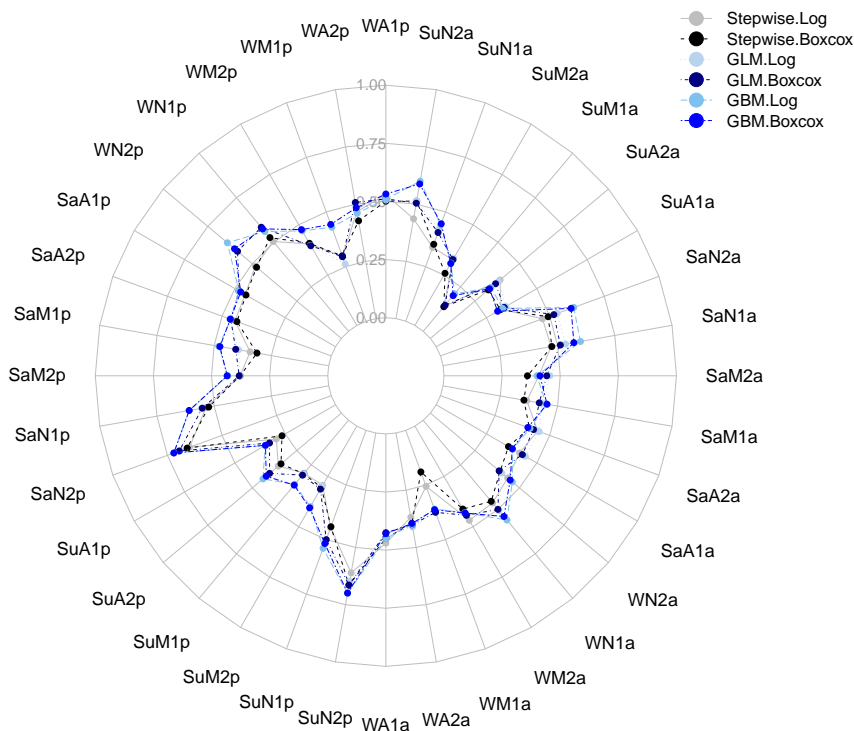


Figure 6.24: Comparison of the R^2 values from the validation of different models

the time intervals. After comparing the 36 time intervals, we can see in Figure 6.25 that the

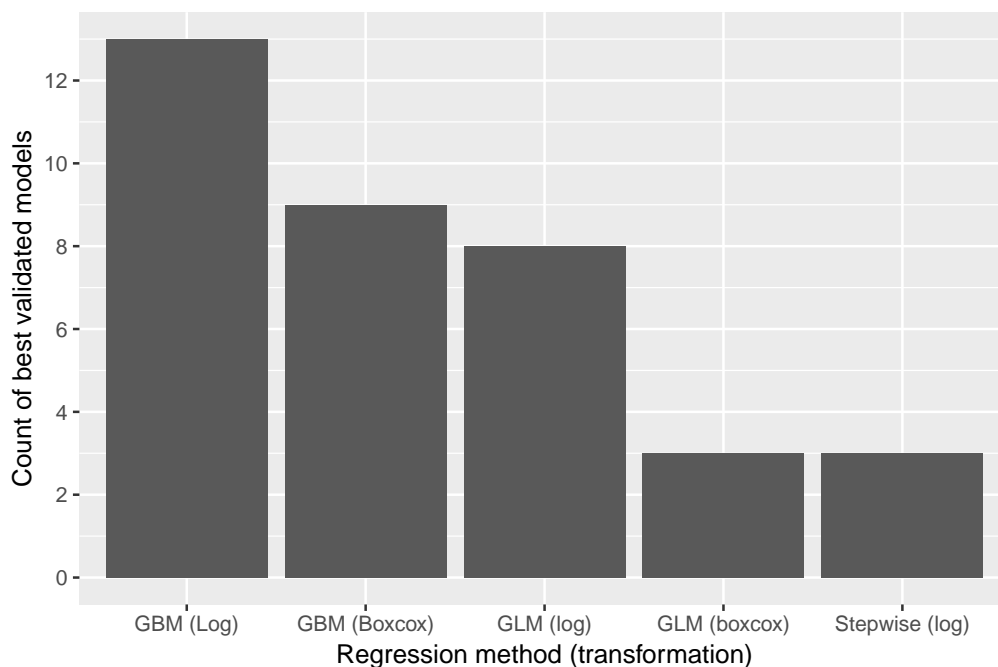


Figure 6.25: Best validated models per regression method

methods fit better the data depending the time interval. They are dependent on the dataset and not on the time intervals since no correlation was shown between the models methods and

the time intervals. GBM usually presented higher R^2 values from the validation and also that the logarithmic transformation performed better more times than the boxcox transformation. As an average, the R^2 from the validation is 0.47 being the highest of all the outcome models. They are dependent on the dataset since no correlation was shown between the models and the time intervals. It is worth emphasizing that the models that included the original dataset without transformations presented the lowest R^2 in the validation.

However, the selected models were considered to be also as simple as possible and not only to fit the data as bet as possible. Stepwise regression was the method with the least number variables selected in a range from around 10 to 25, but GLM and GBM ranged mainly from around 20 to 50. Figure 6.26 is one of the most important outcomes of the thesis, it shows the relationship between the validation results with the number of selected variables of the 324 models. The models that met both criteria were in the upper left part of the graphic, this means those that presented the highest R^2 from the validation and the lowest number of variables. In this case of study, some models built from stepwise regression with boxcox and logarithmic transformations were the models that best-fitted both selection criteria. Nevertheless, a big step is shown after the stepwise regression with values of R^2 under 0.6. Since GBM and GLM presented similar results, the best method could not be estimated from this plot since the number of variables selected from GLM and GBM are unstable because they dependent on the input arguments set by the author. Finally, the models with the worst performance were those where transformations were not considered. Moreover, there was not a clear difference between

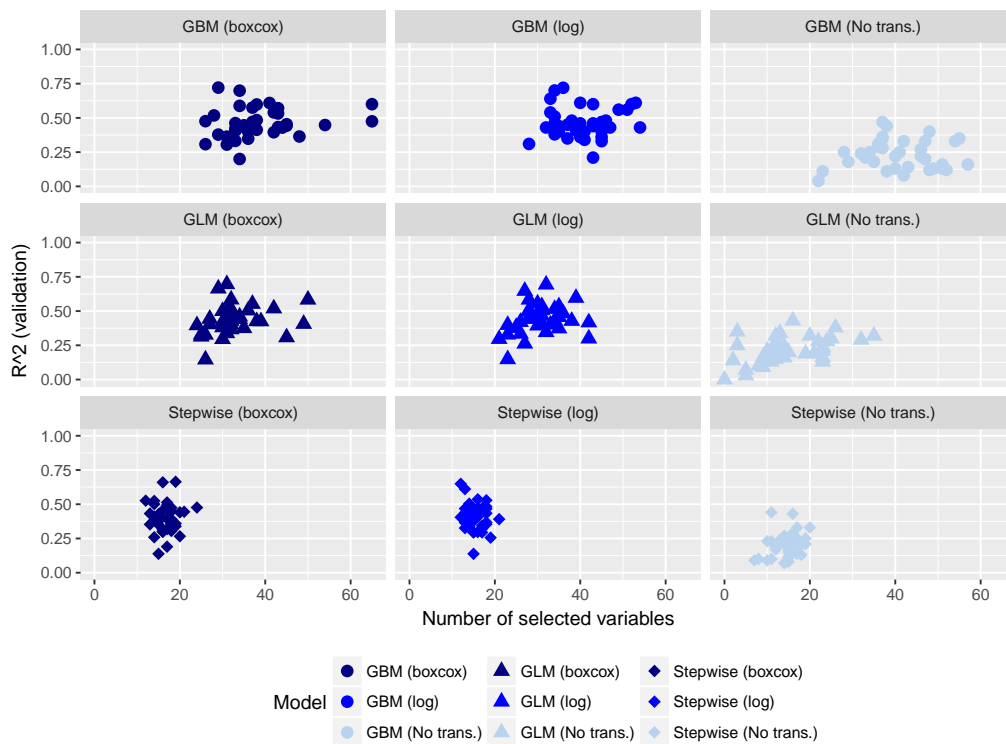


Figure 6.26: Models comparison

logarithmic and boxcox transformation of which met better the criteria. Figure 6.27 shows the Pearson's correlation coefficients between the different variables selected for each model. The highest correlations, as expected, are between the transformations of each method with values ranging from around 0.7 to 1. Commonly, GBM with a logarithmic transformation vary the most in comparison with the other models, followed by GBM with a boxcox transformation.

In conclusion, a regression method could not be selected as "best performer" because neither of the three met both criteria. The three considered regression methods after logarithmic

and boxcox transformations fitted well the data depending on the time interval. However, regarding the better-fitting criteria GBM with a logarithmic transformation presented the better performance and stepwise regression with a logarithmic transformation showed the best results because of selecting lowest number of variables.

It is worth to mention that a overfitting control of the models was not taken into account for the assessment as suggested in the methodological framework, since the models presented similar validation results after the transformations. The highest difference between the R^2_{adj} and the R^2 from the validation was with the GBM method. But this method was also the one that showed the highest validation results.

6.2.4 Principal variables selection

In this section, the variables that influenced the most on the built models are ranked. After the model assessment, we saw that the models with logarithmic and boxcox transformations fitted better the dataset. The variables selected for the six type of models were aggregated per time interval and ranked per total frequency (see Figure 6.28). The most significant variables were: the city population, distance to the city center, bakeries, bicycle parkings, memorials, residential areas, car sharing parking, the density of trees and residential streets and the presence in the zone of influence of pubs, artworks, and universities.

Furthermore, Figure 6.29 and Figure 6.30 show the variables with most relative influence on the GBM with logarithmic and boxcox transformations respectively. This variables are shown because this models presented the better fit with the dataset. The variables are similar than the previous ranking list. It is remarkable the influence of the population and the distance to the city center on the arrivals and departures. A variable that was not high ranked in the previous list was the distance to restaurants. Finally, the variables from the stepwise regression with the logarithmic transformation are also presented in Figure 6.31. This variables are displayed since stepwise regression with logarithmic transformation was the method that presented the fewest number of variables selected with the highest fitted values. Also, the main variables are those included in the previous figures.

6.2.5 Performance of the regression methods using other test cities

As a result of the models' assessment, GBM with a logarithmic transformation predicted mostly better the dataset (see Figure 6.24). An extra analysis was carried out to evaluate the performance of other cities used as test set. Hamburg and Frankfurt were excluded from this analysis because, as mentioned before, they involved the 75% of the observations. Table 6.15 shows the R^2 per time interval from validating the models built after testing on Marburg, Darmstadt and Stuttgart respectively. In other words, as Kassel was used before as test set and the other cities performed as training set, the models were built again using Marburg and then Darmstadt and finally Stuttgart as test sets. Only Stuttgart presented in average a higher values than Kassel with a difference of around 0.07.



Figure 6.27: Correlation between the different methods variables

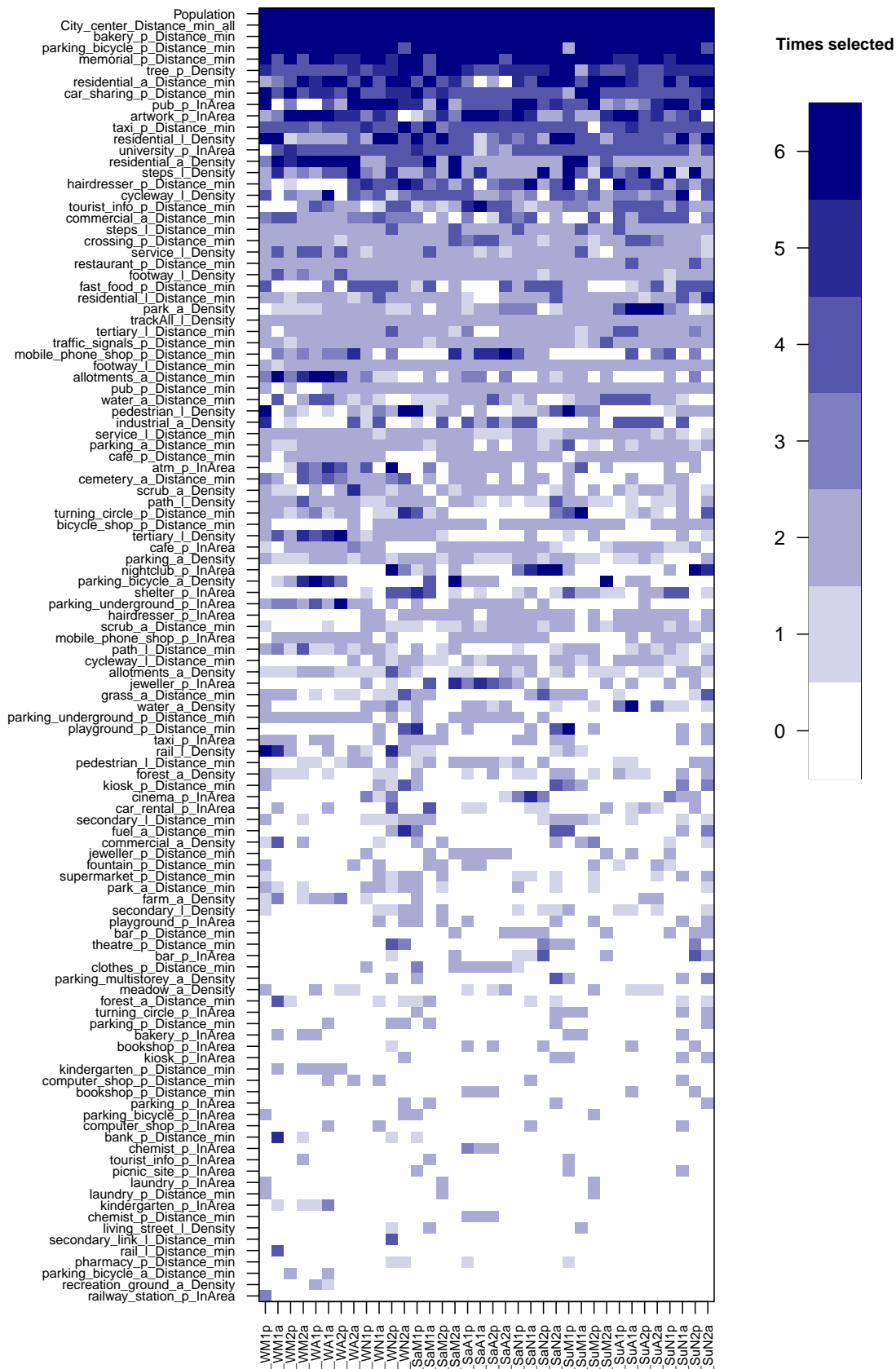


Figure 6.28: Variables with most influence on the built models

Table 6.15: R^2 from validation by testing other cities (GBM with log transformation)

Time	Kassel	Marburg	Darmstadt	Stuttgart
WA1p	0.51	0.39	0.44	0.56
WA2p	0.46	0.39	0.35	0.53
WM1p	0.43	0.32	0.13	0.44
WM2p	0.47	0.30	0.28	0.55
WN1p	0.56	0.45	0.45	0.72
WN2p	0.64	0.36	0.48	0.61
SaA1p	0.48	0.27	0.54	0.43
SaA2p	0.46	0.27	0.49	0.54
SaM1p	0.48	0.06	0.28	0.55
SaM2p	0.43	0.35	0.32	0.38
SaN1p	0.61	0.20	0.43	0.60
SaN2p	0.72	0.49	0.52	0.74
SuA1p	0.35	0.33	0.46	0.34
SuA2p	0.44	0.37	0.45	0.40
SuM1p	0.36	0.12	0.31	0.38
SuM2p	0.40	0.32	0.32	0.31
SuN1p	0.54	0.38	0.35	0.69
SuN2p	0.70	0.52	0.60	0.73
WA1a	0.45	0.31	0.37	0.60
WA2a	0.40	0.42	0.32	0.60
WM1a	0.36	0.51	0.31	0.57
WM2a	0.43	0.46	0.40	0.58
WN1a	0.56	0.32	0.44	0.58
WN2a	0.46	0.19	0.25	0.53
SaA1a	0.38	0.34	0.50	0.57
SaA2a	0.41	0.35	0.54	0.55
SaM1a	0.45	0.35	0.39	0.62
SaM2a	0.40	0.29	0.51	0.67
SaN1a	0.60	0.33	0.51	0.65
SaN2a	0.61	0.26	0.37	0.50
SuA1a	0.33	0.32	0.53	0.48
SuA2a	0.34	0.38	0.41	0.49
SuM1a	0.21	0.18	0.23	0.17
SuM2a	0.31	0.21	0.42	0.57
SuN1a	0.44	0.15	0.45	0.54
SuN2a	0.60	0.13	0.45	0.57
Average	0.47	0.32	0.41	0.54

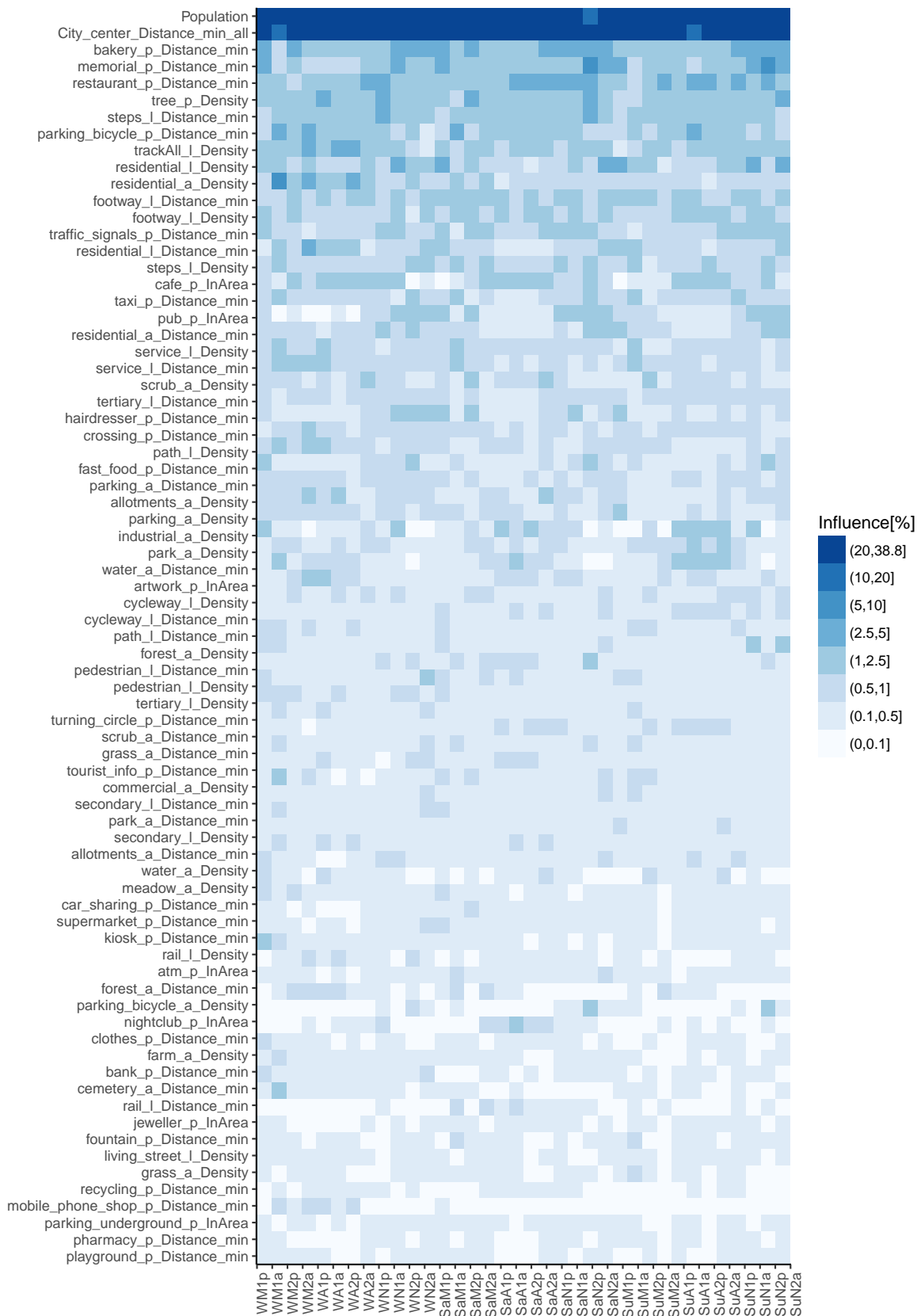


Figure 6.29: Variables with most influence (GBM Logarithmic transformation)

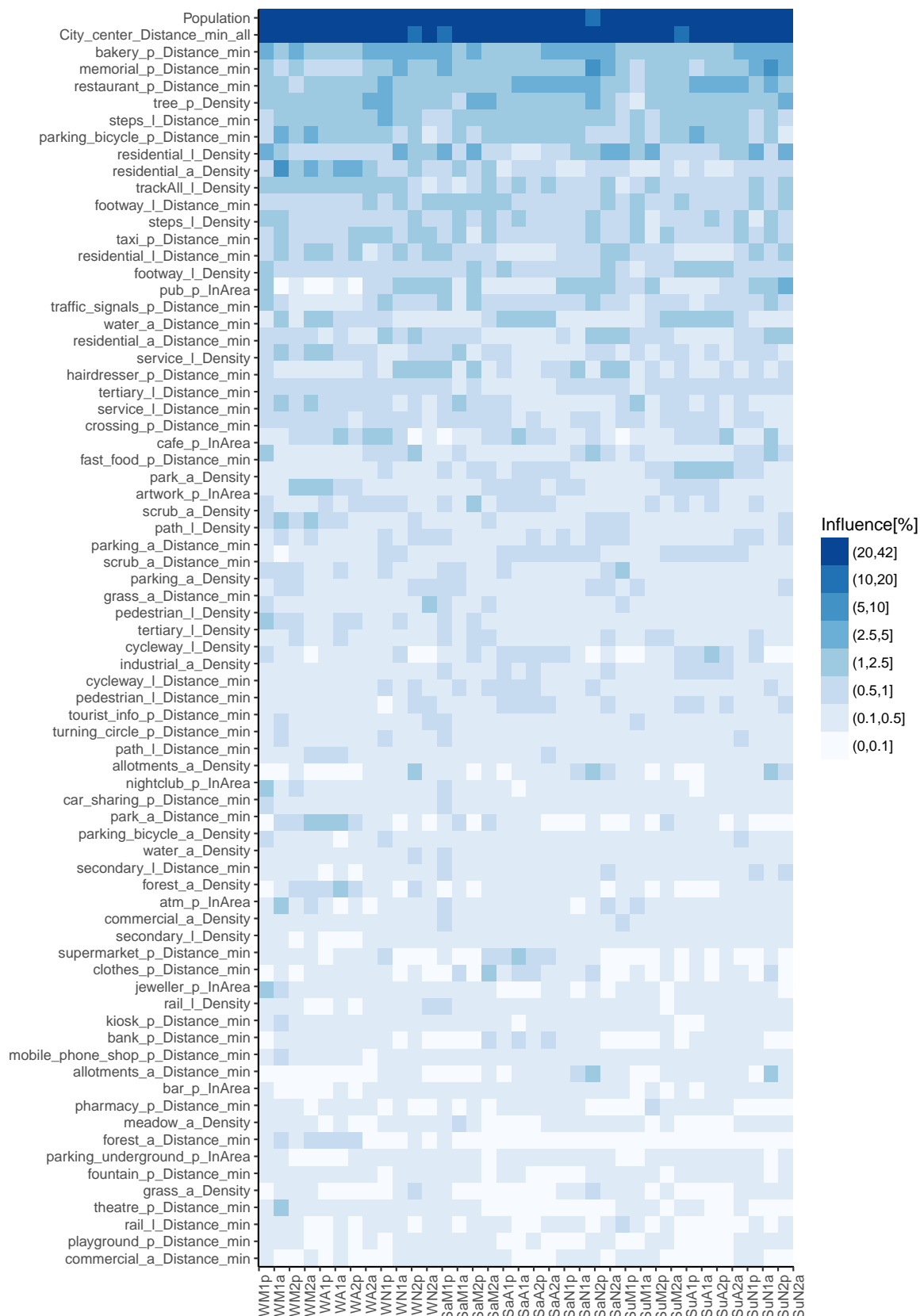


Figure 6.30: Variables with most influence (GBM Boxcox transformation)

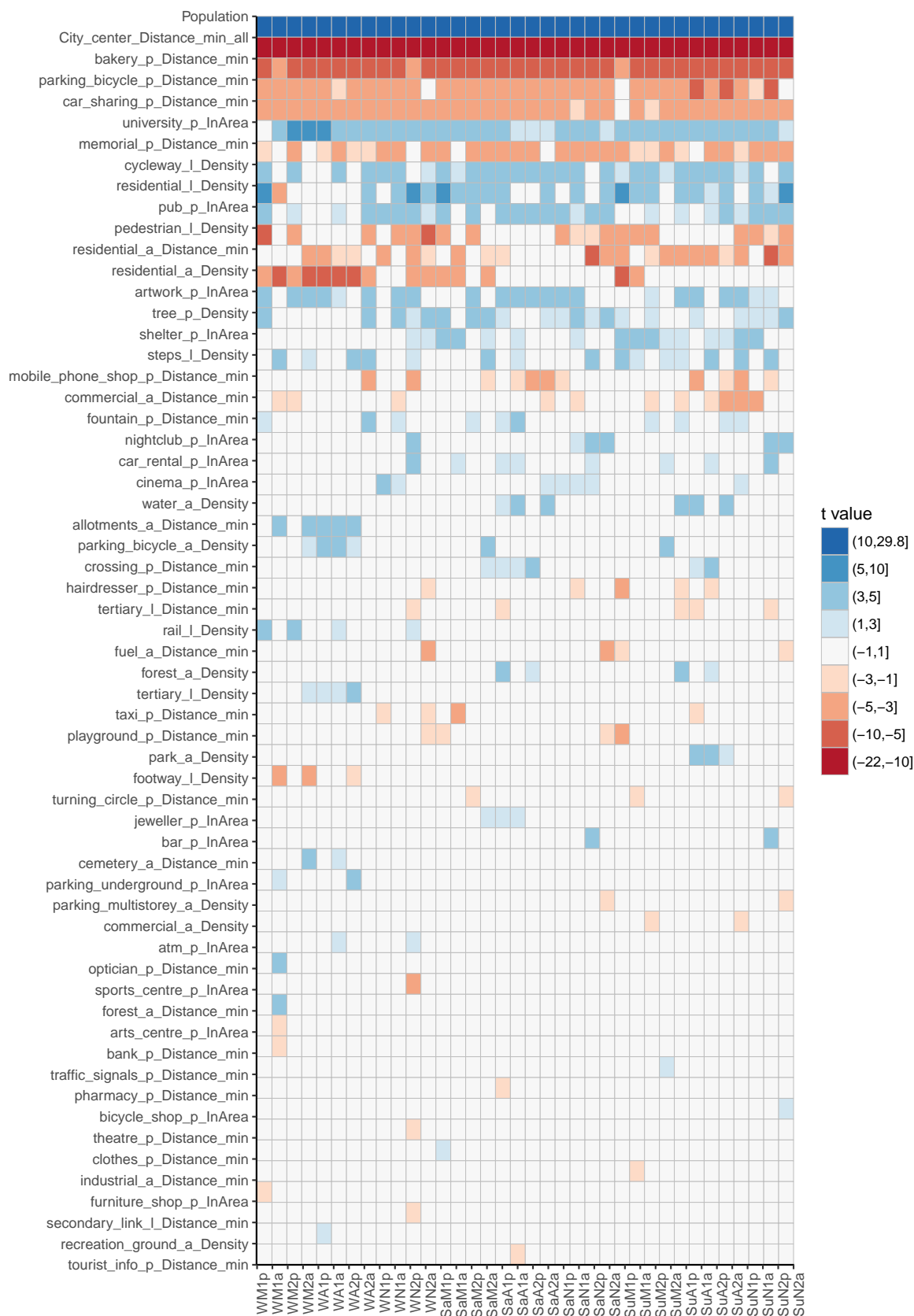


Figure 6.31: Variables with most influence (Stepwise regression Logarithmic transformation)

6.3 Discussion of the results

Arrivals and departures were correlated with exogenous factors in six cities in Germany using the bike sharing system: "Call a bike". The data were collected and analyzed to build 324 models using three regression methods. In this section, the obtained results are discussed starting from the analysis of the collected data and their spatial and temporal units, followed by the selection of models and principal variables.

Five cities were included in the training set and one in the test set. From all the cities, Hamburg had the highest demand and also the highest rate of trips per thousand inhabitants, showing the best performance of the bike sharing systems in the six cities. Frankfurt was the city with the most stations but one of the lowest rates of demand per thousand inhabitants. However, its demand displayed a relatively steep growing trend, so its performance might improve in the future. In Kassel, Konrad displayed relatively high travel times and travel distances, and a high rate of trips per thousand inhabitants, but it was the only city to show that demand is decreasing. Stuttgart presented the lowest rate of trips per thousand inhabitants, the shortest travel times, and one of the lowest travel distances even though it is not one of the smaller cities in the study. The rentals in this city have stayed constant in time. A possible cause of the low usage rate is hilly the topography, the higher cost of the electric bikes, and also that the stations were less densely placed than in the other systems. Moreover, Marburg and Darmstadt have both shown high performance and higher growth. These cities provide proof that giving benefits to students to join BS systems can achieve good results.

According to the assignment of indicators to the exogenous factors, logic indicators were adopted. For example, tree density was considered instead of the presence of a tree in an area, and bank presence instead of bank density. However, for the variables categorized as spatial points, most of them used the distance and presence inside each zone of influence. The indicators of density of spatial points and distance to points from a station within the city were almost not present after the preselection phase. They were either not included in entire cities or they were collinear to other indicators.

When the dataset was processed, the models were built. In the literature review for station-based bike sharing (see Table 2.4), linear regression methods were commonly used. In addition to OLS, GLM and GBM were also used as regression methods. Willing et al. (2017) correlated points of interest with the demand on a car sharing system in Amsterdam using GLM as a regression method and GBM as a variable selection technique. However, in this thesis, a different approach was taken by using both as regression and variable selection techniques. GBM was the regression method that best fit the data, followed by GLM. However, they were more complex than the linear regression after a stepwise regression. It is worth it to mention that GLM presented the lowest computational effort followed by GBM. An advantage of stepwise regression is that only a variable selection technique (AIC or BIC) was required as input argument. The other two models needed cross-validation tests to select the input arguments that helped to build models to better fit the dataset.

Moreover, logarithmic and boxcox transformations were applied to the dependent variables of the three implemented regression methods. Models without the transformation presented illogical variable selection, higher errors, and worse prediction results. The logarithmic transformation was considered as in several cases in the literature (see Table 2.4), however boxcox was usually not present. Generally, for the three regression methods, the logarithmic transformation performed better in most cases with a higher BIC and higher R^2 in the validation phase. However, boxcox presented higher results of the R^2 , R^2_{adj} and MSEs on the fitting phase and lower MSEs on the validation phase. Therefore, it was not possible to decide which transformation was better for the dataset. However, a higher weight was given to the R^2 in the validation phase and therefore the logarithmic transformation was selected as the transformation technique that better validated the data.

The models with time intervals with relative less demand spread between cities presented higher fitting. More than the quantity of the rentals, the spread of the data between cities (see Figure 6.9). For example, for "Workday Night II", where all cities showed similar behavior, the models presented the highest validation values, but mornings on workdays performed worse when the spread of the relative demand between cities was higher. After this fact, Mondays should not have been taken to represent the dependent variables for workdays since Thursdays and Fridays exhibited less spread.

According to the variable selection, stepwise regression method with a BIC selection technique selected variables from a range of 10 to 20. The other two models had a wider range but better validation results. The advantage of stepwise regression and GLM was that a variable selection process was implicit in the methods, but for GBM a variable selection script had to be constructed to select those with more influence from the ranking list. The relative high correlation between the variables selected for each stepwise regression and GLM shows that the methods of direct elimination to deal with collinearity worked.

The most important variables in all of the built models and through all time intervals were the population of the city and the distance from the city center (old town) to the stations. The population of the city helped to weight the models to have a common scale that was not biased if the city was a large like Hamburg or a small city like Marburg. The distance to the city center plays a significant role on the demand as seen in Figure 6.10. The third most influencing variable is the distance to bakeries. If a station is close to a bakery, this increases the probability of a higher demand at that station.

In general, the principal variables are related to leisure activities, parks, green areas, and water bodies on the weekends, especially pubs, cinemas, clubs at night, shops on Saturdays, and memorials outside of working hours. Just few transport-related variables significantly influenced the models. Distance to a car sharing station was significant for all time intervals. This might lead to a correlation between car sharing and bike sharing demand. Another variable that was commonly present was bicycle parking. This fact is obvious because bike sharing stations are usually close to additional bicycle parking. The only public transport variable displayed was railway stations, during workday mornings. It is worth it to say that tram and metro variables were not considered because they were not present in most of the cities.

Additionally, to the number of variables, another important criterion is the logic under the influence of the variables. After aggregation, the instances in which the different models with logarithmic and boxcox transformation selected the variables per time interval showed logical consistencies in the variables selected. For instance, to name a few, the relevance of:

- Park areas, playground, areas with bodies of water and areas with grass on Sundays,
- Railways on workday mornings,
- Night clubs, cinemas, gas stations at night,
- Banks on workdays mornings,
- Pubs at night, but not relevant on working day mornings,
- Universities during working hours,
- Restaurants on Saturday afternoon and night
- Density of cycleways specially out of the working hours showing the importance of the bicycle infrastructure,

Practically speaking, the arrivals and departures were highly correlated with each other with some differences. Also, the predicted models presented similar assessment values. However, for the variables' influence on the models there were remarkable differences. For example, in the

morning, the presence of universities or commercial areas, or the distance to clothing shops were relevant for attractions but not for productions.

Stepwise regression with a logarithmic transformation presented the advantage that the results are easier to understand than GBM, for example. From the sign of the t value, we can see directly if a variable positively or negatively affects the demand. Commonly, the presence in the zone and the density had positive values, while the distance had negative values.

However, not all the variables were logical. Some variables without importance influenced the models. Some examples are sports centers in an area at "Night II", the density of public steps, and presence of furniture shops. A controversial finding is that the residential areas reduced the demand but the density of residential streets increased the demand. Nevertheless, an indicator of the success of the variable selection method used is that the presence of all existing infrastructures were categorized as factors influencing the demand in SBBS systems, as seen in the literature review (see Table 2.5).

Finally, workday mornings presented the lowest R^2 values from the validation and also the lowest number of variables selected. The possible reason of this fact might be the different behavior between cities during the morning. Therefore, as a conclusion, mornings should not be used when comparing multiple cities of varying size because they present different relative demand during this time.

Chapter 7

Conclusions

In this last chapter, first, the conclusions of the thesis are presented. Then, recommendations for the implementation of the methodology are discussed, followed by the recommended future work including aspects omitted in the case of study of the thesis.

7.1 Conclusions

Car sharing and bike sharing are two categories of shared mobility, which are defined as a short-term rental of a car or a bicycle respectively, avoiding the costs and responsibilities of owning these vehicles. Their main benefit is the decrease in private cars ownership leading to the reduction of VKTs, efficient use of the roads infrastructure, and trips shift from private cars to active modes. Consequently, the shared mobility concept has usually presented individual economic saving, fewer emissions and health benefits. Because of these benefits, an enhancement of their demand is needed to decrease the harmful effects of private cars usage to the environment and therefore, the society.

Identifying exogenous factors from existing car and bike sharing systems are required to expand such concept to new regions and cities, to increase the reliability of the implementations and to reduce their risk of supply-demand imbalance. In the literature, some authors have identified factors affecting the demand for shared vehicles by usually identifying relationships between them on the logarithm of arrivals and departures rates. The most common method used was linear regression models. The most significant exogenous factor were city size, jobs, and population density, public transport stations, universities, residential land use, distance to the CBD, among others.

Moreover, ICT development has helped to increase significantly the obtaining of data, for instance, in the transport sector and also high precision data on geographical information. In the literature, to the best of the authors' knowledge, there is not a consistent methodology to correlate open-source arrivals and departure rates from shared transportation systems with exogenous factors from open geographic sources in multiple cities in a local scale.

This thesis developed a automated methodology in three main directions: 1) automated data collection from open-source data, 2) automated data analysis and processing and 3) automated model building and selection using three methods: stepwise regression, GLM, and GBM. As a case of study, daily average arrivals and departures from five cities in Germany (Hamburg, Frankfurt, Stuttgart, Marburg, and Darmstadt) using the SBBS system "Call a Bike" were used to automatically identify the relationships with the population and the existing infrastructure obtained from OpenStreetMap. In total 324 models were built as a result of a training set of 631 stations aggregated in 36 times intervals including the departures and arrivals correlated with around 200 independent variables. As heteroscedasticity was identified in the model's outputs, logarithmic and boxcox transformations were implemented to the dependent variables showing a significant improvement of the performance on the validation carried out with 58 zones of

influence from the city of Kassel.

In general, the fitting and validation results from each regression method and transformation technique was dependent on time intervals. However, GBM with a logarithmic transformation of the dependent variables was found to perform better in the most cases of the validation set. In addition to the models building the variables that influence the most the models were selected and ranked. Stepwise regression with a logarithmic transformation of the dependent variables was found to perform adequately with fewer variables than other models. However, the validation results were the lowest of the three methods. The most influencing variables selected were the city population, the distance from the stations to the city center, bakeries, memorials, car sharing stations, among others. Logical relationships between the variables with the time intervals were displayed, such as higher demand on nights close to pubs, cinemas, and nightclubs; or the presence of water bodies, parks or green areas on Sundays. Another indicator of the good performance of the variables selection is the fact that the chosen variables were present in diverse researchers on the literature review.

The models that fitted better the data were in the afternoon and at night where the different cities showed a more common hourly relative demand. At these times, models presented better results from the validation, and performed a more logical selection of variables that influenced the demand. Therefore, to obtain models that fit better the data is recommended to build them by selecting multiple cities with relative temporal patterns that are similar.

7.2 Recommendations and future work

This study developed a consistent methodology to correlate arrivals and departures of shared vehicles in multiple cities on a local level using open-source data and also rank the variables that influence the most. To implement the methodology of this thesis is recommended not to use original arrivals and departures rates but transformations using a concave function as the logarithm. Depending on the purpose of the implementation a different model is recommended to be implemented. For example, to predict the demand GBM models are more recommended but to select the variables that influence the most stepwise regression is recommended. Also, is recommended to cluster the studied cities according to the temporal distribution of their relative demand.

From this thesis methodology and results, further research and implementation can be developed as improvements and expansion of the case of study, correlation analysis between car and bike sharing, identifying exogenous factors affecting the trips distribution of shared vehicles, and practical implementations of the developed methodology and case of study. The possible further work is presented below.

Improvements in the case study. After the relative good validation of the models and the logical influencing variables, further improvement of the case of study can be performed as it follows.

To collect a better database including the FFBS systems of "Call a Bike" in Munich or Cologne. Then, further analysis would consist to combine SBBS and FFBS in the study.

- To consider temporal variations of the exogenous factors.
- To include exogenous factors that were not considered, such as the topography and population and jobs density, mobility behavior, socio-demographic factors, among others.
- To extend the analysis of the study by clustering the cities according to the number of inhabitants or same hourly demand patterns.
- To increase temporal precision considering the months in the time units. For instance: April workday morning peak time

Expansion of the approach of the case of study. The thesis can be further implemented by including more cities in the study in a multinational level and analyze if the variables are concurrent at an international level. Also, the methodology of this thesis can be expanded to the analysis of other shared vehicles car sharing systems or free-floating based systems.

Expansion of the approach of the methodology. In the methodological framework just three regression methods were test to identify the relationships between the dependent and the independent variables. However, other regression methods can be tested to analyze if they fit better the dataset or perform a better selection of the most influencing variables.

Correlations between car and bike sharing. An extra approach would be to consider the demand of car sharing as an independent variable of bike sharing and vice versa, as the outcome of this thesis showed the relationship between car and bike sharing. Additionally, there was found a need to treat bike and car sharing together as part of a shared mobility system. This requirement is to understand the relationships between them, i.e., if they conflict or they support each other.

Identifying exogenous factors affecting the trips distribution of shared vehicles. The approach from trips origin and destinations can be extended to trips distribution. In other words, to study if the frequency of routes between stations is affected by exogenous factors.

Practical implementation This thesis will help to enhance shared mobility, by showing the validity and increase the reliability of measures, policies, and shared mobility projects. For example, it can be implemented to analyze the location of stations based on the land-use. The models were carried out in the cities in Germany with the highest demand, so the built models can give an idea to identify what causes the lower demand in the other cities. Also, the BS systems can be improved by, for example, manipulating the indicators to promote the change of land-use for a better development of the system. For instance, changes in the activities close to the stations can be promoted as bakeries for instance. Finally, the success factors of shared mobility can be extended to other cities of the region or the world, especially to shape shared mobility in developing cities where sustainable mobility is starting to grow.

Bibliography

- Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265.
- Banister, D. (2008). The sustainable mobility paradigm. *Transport Policy*, 15(2):73 – 80. New Developments in Urban Transportation Planning.
- Bishara, A. and Hittner, J. (2012). Testing the significance of a correlation with nonnormal data: comparison of pearson, spearman, transformation, and resampling approaches. *Psychological methods*, 17(3):399.
- Böckmann, M. (2013). The shared economy: It is time to start caring about sharing; value creating factors in the shared economy. *University of Twente, Faculty of Management and Governance*.
- Böhm, O. (2016). Stadtrad hamburg. <https://www.klimaaktiv.at/dam>. Accessed on: 31.10.2017.
- Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252.
- Büttner, J. and Petersen, T. (2011). *Optimising Bike Sharing in European Cities: A Handbook*. OBIS.
- CarSharing eV Bundesverband (2017a). Geschichte. <https://carsharing.de/alles-ueber-carsharing/ist-carsharing/geschichte>. Accessed on: 12.10.2017.
- CarSharing eV Bundesverband (2017b). Unterschiede free-floating & stationsbasiertes carsharing. <https://carsharing.de/presse/fotos/zahlen-daten/unterschiede-free-floating-stationsbasiertes-carsharing>. Accessed on: 11.10.2017.
- Carsharing News (2017). Carsharing anbieter. www.carsharing-news.de/carsharing/. Accessed on: 11.10.2017.
- Caulfield, B., O'Mahony, M., Brazil, W., and Weldon, P. (2017). Examining usage patterns of a bike-sharing scheme in a medium sized city. *Transportation Research Part A: Policy and Practice*, 100:152 – 161.
- Celsor, C. and Millard-Ball, A. (2007). Where does carsharing work?: Using geographic information systems to assess market potential. *Transportation Research Record: Journal of the Transportation Research Board*, pages 61–69.
- Chardon, C. M. D., Caruso, G., and Thomas, I. (2017). Bicycle sharing system ‘success’ determinants. *Transportation Research Part A: Policy and Practice*, 100:202 – 214.
- Chatterjee, S. and Hadi, A. (2015). *Regression analysis by example*. John Wiley and Sons.

- Cheng, M. (2016). Sharing economy: A review and agenda for future research. *International Journal of Hospitality Management*, 57:60 – 70.
- Clark, J. and Curl, A. (2016). Bicycle and car share schemes as inclusive modes of travel? a socio-spatial analysis in glasgow, uk. *Social Inclusion*, 4(3):83–99.
- Cohen, B. and Muñoz, P. (2016). Sharing cities and sustainable consumption and production: towards an integrated framework. *Journal of Cleaner Production*, 134, Part A:87 – 97. Special Volume: Transitions to Sustainable Consumption and Production in Cities.
- Comendador, J., Lopez-Lambas, M. E., and Monzon, A. (2014). Urban built environment analysis: Evidence from a mobility survey in madrid. *Procedia - Social and Behavioral Sciences*, 160:362 – 371.
- DB Rent GmbH (2015). Call a bike – stadt frankfurt. <http://bikeandbusiness.de>. Accessed on: 12.10.2017.
- Deutsche Bahn AG (2017). Das smarte leihfahrrad der deutschen bahn — call a bike. <https://www.callabike-interaktiv.de/de>. Accessed on: 31.10.2017.
- Deutsche Bahn (DB) (2017). Call a bike - open-data-portal – deutsche bahn datenportal. <http://data.deutschebahn.com/dataset/data-call-a-bike>. Accessed on: 31.10.2017.
- Deutsches Institut für Urbanistik (2011). Fahrrad- und e-bike-verleihsystem für stuttgart. <https://nationaler-radverkehrsplan.de/de/praxis/fahrrad-und-e-bike-verleihsystem-fuer-stuttgart>. Accessed on: 12.10.2017.
- Deutsches Institut für Urbanistik (2015). Nationale und internationale entwicklungen oeffentliche fahrradverleihsysteme. <https://nationaler-radverkehrsplan.de/de/forschung/schwerpunktthemen/oeffentliche-fahrradverleihsysteme>. Accessed on: 11.10.2017.
- EC (2016). Statistical pocketbook 2016 - mobility and transport. Technical report, European Commission.
- Efthymiou, D., Antoniou, C., and Waddell, P. (2013). Factors affecting the adoption of vehicle sharing systems by young drivers. *Transport Policy*, 29:64 – 73.
- El-Assi, W., Mahmoud, M., and Habib, K. (2017). Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in toronto. *Transportation*, 44(3):589–613.
- Encyclopaedia Britannica. Frankfurt am main. www.britannica.com/place/Frankfurt-am-Main.
- Encyclopaedia Britannica. Hamburg. <https://www.britannica.com/place/Hamburg-Germany>. Accessed on : 31.10.2017.
- Encyclopaedia Britannica. Stuttgart. [/www.britannica.com/place/Stuttgart-Germany](http://www.britannica.com/place/Stuttgart-Germany).
- Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.
- Faghieh-Imani, A. and Eluru, N. (2016). Incorporating the impact of spatio-temporal interactions on bicycle sharing system demand: A case study of new york citibike system. *Journal of Transport Geography*, 54:218 – 227.

- Faghih-Imani, A., Eluru, N., El-Geneidy, A. M., Rabbat, M., and Haq, U. (2014). How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (bixi) in montreal. *Journal of Transport Geography*, 41:306 – 314.
- Faghih-Imani, A., Hampshire, R., Marla, L., and Eluru, N. (2017). An empirical analysis of bike sharing usage and rebalancing: Evidence from barcelona and seville. *Transportation Research Part A: Policy and Practice*, 97(Supplement C):177 – 191.
- Firnkorn, J. (2012). Triangulation of two methods measuring the impacts of a free-floating carsharing system in germany. *Transportation Research Part A: Policy and Practice*, 46(10):1654–1672.
- Firnkorn, J. and Müller, M. (2011). What will be the environmental effects of new free-floating car-sharing systems? the case of car2go in ulm. *Ecological Economics*, 70(8):1519–1528.
- Firnkorn, J. and Shaheen, S. (2016). Generic time- and method-interdependencies of empirical impact-measurements: A generalizable model of adaptation-processes of carsharing-users' mobility-behavior over time. *Journal of Cleaner Production*, 113:897 – 909.
- Fishman, E., Washington, S., and Haworth, N. (2013). Bike share: A synthesis of the literature. *Transport Reviews*, 33(2):148–165. cited By 98.
- Fishman, E., Washington, S., and Haworth, N. (2014). Bike share's impact on car use: evidence from the united states, great britain, and australia. *Transportation Research Part D: Transport and Environment*, 31:13–20.
- Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232. cited By 2353.
- Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.
- Gauthier, A., Hughes, C., Kost, C., Li, S., et al. (2013). The bike-share planning guide. itdp report. Accessed on: 31.10.2017.
- Gießener Anzeiger Verlags GmbH (2017). Call a bike: Marburger asta organisiert leihraeder. <http://www.giessener-anzeiger.de>. Accessed on: 31.10.2017.
- Giesel, F. and Nobis, C. (2016). The impact of carsharing on car ownership in german cities. *Transportation Research Procedia*, 19:215 – 224. Transforming Urban Mobility. mobil.TUM 2016. International Scientific Conference on Mobility and Transport. Conference Proceedings.
- Haklay, M. (2010). How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, 37(4):682–703.
- HAMBURG.DE (2017). Hamburg in zahlen. <http://www.hamburg.de>. Accessed on: 31.10.2107.
- Hauke, J. and Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.
- Janssen, M., Charalabidis, Y., and Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management*, 29(4):258–268.

- Kang, J., Hwang, K., and Park, S. (2016). Finding factors that influence carsharing usage: Case study in seoul. *Sustainability*, 8:709.
- Kent, J. L. (2014). Carsharing as active transport: What are the potential health benefits? *Journal of Transport and Health*, 1(1):54–62.
- Kopp, J., Gerike, R., and Axhausen, K. W. (2015). Do sharing people behave differently? an empirical evaluation of the distinctive mobility patterns of free-floating car-sharing members. *Transportation*, 42(3):449–469.
- Kortum, K., Schönduwe, R., Stolte, B., and Bock, B. (2016). Free-floating carsharing: City-specific growth rates and success factors. *Transportation Research Procedia*, 19:328 – 340.
- Lin, D., Foster, D. P., and Ungar, L. H. (2011). Vif regression: A fast regression algorithm for large data. *Journal of the American Statistical Association*, 106(493):232–247.
- Mack, C. (2016). Lecture notes in from data to decisions: Measurement, uncertainty, analysis, and modeling. <http://www.lithoguru.com/scientist/statistics/>. Accessed on: 11.10.2017.
- Mackett, R. L. and Thoreau, R. (2015). Transport, social exclusion and health. *Journal of Transport & Health*, 2(4):610 – 617.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297. University of California Press.
- Martin, E., Shaheen, S., and Lidicker, J. (2010). Impact of carsharing on household vehicle holdings. *Transportation Research Record*, (2143):150–158. cited By 94.
- Martin, E. W. and Shaheen, S. A. (2011). Greenhouse gas emission impacts of carsharing in north america. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1074–1086.
- Mattson, J. and Godavarthy, R. (2017). Bike share in fargo, north dakota: Keys to success and factors affecting ridership. *Sustainable Cities and Society*, 34:174 – 182.
- Meddin, R. and DeMaio, P. (2015). The bike-sharing world map. <http://www.metrobike.net>.
- Miller, A. J. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)*, 147(3):389–425.
- Noland, R. B., Smart, M. J., and Guo, Z. (2016). Bikeshare trip generation in new york city. *Transportation Research Part A: Policy and Practice*, 94:164 – 181.
- Nowicka, K. (2016). Cloud computing in sustainable mobility. *Transportation Research Procedia*, 14:4070 – 4079. Transport Research Arena TRA2016.
- OpenStreetMap-contributors (2017). Planet dump retrieved from <https://planet.osm.org> . <https://www.openstreetmap.org> . Accessed on: 31.10.2017.
- Pal, A. and Zhang, Y. (2017). Free-floating bike sharing: Solving real-life large-scale static rebalancing problems. *Transportation Research Part C: Emerging Technologies*, 80:92 – 116.
- Posada, D. and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793–808.
- Reiss, S. and Bogenberger, K. (2017). A relocation strategy for munich's bike sharing system: Combining an operator-based and a user-based scheme. *Transportation Research Procedia*, 22:105–114.

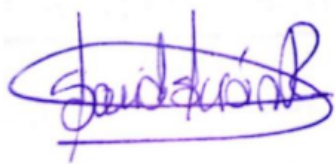
- Ricci, M. (2015). Bike sharing: A review of evidence on impacts and processes of implementation and operation. *Research in Transportation Business & Management*, 15:28 – 38. Managing the Business of Cycling.
- Ridgeway, G. (2006). gbm: Generalized boosted regression models. *R package version*, 1(3):55.
- Rodrigue, J.-P., Comtois, C., and Slack, B. (2016). *The geography of transport systems*. Taylor & Francis.
- Schmöller, S. and Bogenberger, K. (2014). Analyzing external factors on the spatial and temporal demand of car sharing systems. *Procedia - Social and Behavioral Sciences*, 111:8 – 17.
- Schmöller, S., Weikl, S., Müller, J., and Bogenberger, K. (2015). Empirical analysis of free-floating carsharing usage: The munich and berlin case. *Transportation Research Part C: Emerging Technologies*, 56:34 – 51.
- Schubert, A. (2017). Obikes in münchen: Imageschaden im eiltempo. <http://www.sueddeutsche.de/muenchen>. Accessed on: 12.10.2017.
- Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.
- Seign, R., Schüßler, M., and Bogenberger, K. (2015). Enabling sustainable transportation: The model-based determination of business/operating areas of free-floating carsharing systems. *Research in Transportation Economics*, 51:104 – 114. Austerity and Sustainable Transportation.
- Shaheen, S. and Chan, N. (2015). Mobility and the sharing economy: Impacts synopsis. *Transportation Sustainability Research Center, University of California, Berkeley*. http://tsrc.berkeley.edu/sites/default/files/Innovative-Mobility-Industry-Outlook_SM-Spring-2015.pdf.
- Shaheen, S., Chan, N., Bansal, A., and Cohen, A. (2015). Shared mobility: Definitions, industry development, and early understanding. *Transportation Sustainability Research Center (TSRC), UC Berkeley*.
- Shaheen, S. and Cohen, A. (2016). Innovative mobility carsharing outlook. transportation sustainability research center. *University of California, Berkeley*. Retrieved December, 19:2016.
- Shaheen, S., Guzman, S., and Zhang, H. (2010a). Bikesharing in europe, the americas, and asia: past, present, and future. *Transportation Research Record: Journal of the Transportation Research Board*, (2143):159–167.
- Shaheen, S., Martin, E., Cohen, A., and Finson, R. (2012). Public bikesharing in north america: Early operator and user understanding, mti report 11-19. Technical report, Mineta Transportation Institute.
- Shaheen, S., Rodier, C., Murray, G., Cohen, A., and Martin, E. (2010b). Carsharing and public parking policies: assessing benefits, costs, and best practices in north america. Technical report, Mineta Transportation Institute.
- Shaheen, S. A. (2016). Mobility and the sharing economy. *Transport Policy*, 51:141 – 142. Transit Investment and Land Development. Edited by Xinyu (Jason) Cao and Qisheng Pan & Shared Use Mobility Innovations. Edited by Susan Shaheen.
- Shaheen, S. A. and Cohen, A. P. (2012). Carsharing and personal vehicle services: Worldwide market developments and emerging trends. *International Journal of Sustainable Transportation*, 7(1):5–34.

- Stadt Frankfurt am Main (2017). Bevoelkerung. www.frankfurt.de.
- Stadt Kassel (2016). Fahrradvermietsystem konrad. <http://www.kassel.de/stadt/mobilitaet>. Accessed on: 12.10.2017.
- Stadt Kassel (2017). Stadtinformation. <http://www.kassel.de/stadt/>. Accessed on: 31.10.2017.
- Statista (2016). Entwicklung der Einwohnerzahl in Stuttgart (kreisfreie Stadt) von 1995 bis 2015. de.statista.com. Accessed on: 31.10.2017.
- Statistisches Bundesamt (2012). Bevölkerung und erwerbstätigkeit. bevölkerung mit migrationshintergrund. ergebnisse des mikrozensus 2011. fachserie 1, reihe 2.2. <https://www.destatis.de/DE/Publikationen/>. Accessed on: 31.10.2017.
- Statistisches Bundesamt (2016). Ein sozialbericht für die bundesrepublik deutschland. *Bonn: Bundeszentrale für politische Bildung*.
- Suche-postleitzahl.org (2017). Postleitzahlen deutschland • plz suche deutschland. <http://www.suche-postleitzahl.org>. Accessed on: 31.10.2017.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.
- Tran, T. D., Ovtracht, N., and d’Arcier, B. F. (2015). Modeling bike sharing system using built environment factors. *Procedia CIRP*, 30:293 – 298. 7th Industrial Product-Service Systems Conference - PSS, industry transformation for sustainability and business.
- Tyndall, J. (2017). Where no cars go: Free-floating carshare and inequality of access. *International Journal of Sustainable Transportation*, 11(6):433–442.
- Universitätsstadt Marburg (2017). Lage, struktur und daten. <https://www.marburg.de/wirtschaft-universitaet/stadt-region-und-wirtschaft/stadt-und-region/lage-struktur-und-daten/>. Accessed on: 12.10.2017.
- Voronoi, G. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire. recherches sur les paralléloèdres primitifs. *Journal für die reine und angewandte Mathematik*, 134:198–287.
- Wang, X., Lindsey, G., Schoner, J., and Harrison, A. (2015). Modeling bike share station activity: Effects of nearby businesses and jobs on trips to and from stations. *Journal of Urban Planning and Development*, 142(1):04015001.
- Weikl, S. and Bogenberger, K. (2013). Relocation strategies and algorithms for free-floating car sharing systems. *IEEE Intelligent Transportation Systems Magazine*, 5(4):100–111.
- Willing, C., Klemmer, K., Brandt, T., and Neumann, D. (2017). Moving in time and space – location intelligence for carsharing decision support. *Decision Support Systems*, 99:75 – 85. Location Analytics and Decision Support.
- Wissenschaftsstadt Darmstadt (2017). Stadtporträt. <https://www.darmstadt.de/standort/stadtportraet/>. Accessed on: 12.10.2017.
- Zhao, J., Deng, W., and Song, Y. (2014). Ridership and effectiveness of bikesharing: The effects of urban features and system characteristics on daily use and turnover rate of public bikes in china. *Transport Policy*, 35:253 – 264.
- Zielstra, D. and Zipf, A. (2010). Quantitative studies on the data quality of openstreetmap in germany. In *Proceedings of the Sixth International Conference on Geographic Information Science, GIScience, Zurich, Switzerland*, pages 20–26.

Declaration concerning the Master's Thesis

I hereby confirm that the presented thesis work has been done independently and using only the sources and resources as are listed. This thesis has not previously been submitted elsewhere for purposes of assessment.

Munich, November 1st, 2017

A handwritten signature in blue ink, appearing to read 'David Durán Rodas', enclosed within a blue oval scribble.

David Durán Rodas

Appendices

Appendix A

Summary of the independent variables

Table A.1: Summary of the independent variables

Statistic	Mean	St. Dev.	Min	Median	Max
allotments_a_Density	0.02	0.05	0.00	0.00	0.37
allotments_a_Distance_min	73,931.51	43,866.96	1.35	99,999.00	99,999.00
arts_centre_p_InArea	0.05	0.21	0	0	1
artwork_p_InArea	0.26	0.44	0	0	1
atm_p_InArea	0.31	0.46	0	0	1
attraction_p_Distance_min	93,044.95	25,430.92	20.75	99,999.00	99,999.00
attraction_p_InArea	0.07	0.25	0	0	1
bakery_p_Distance_min	45,494.44	49,765.33	3.97	289.89	99,999.00
bakery_p_InArea	0.55	0.50	0	1	1
bank_p_Distance_min	64,485.19	47,843.33	6.46	99,999.00	99,999.00
bank_p_InArea	0.36	0.48	0	0	1
bar_p_Distance_min	69,569.13	45,991.11	16.76	99,999.00	99,999.00
bar_p_InArea	0.30	0.46	0	0	1
beauty_shop_p_Distance_min	82,607.81	37,897.49	8.54	99,999.00	99,999.00
beauty_shop_p_InArea	0.17	0.38	0	0	1
beverages_p_Distance_min	82,324.93	38,129.73	27.86	99,999.00	99,999.00
beverages_p_InArea	0.18	0.38	0	0	1
bicycle_shop_p_Distance_min	80,872.88	39,317.25	14.14	99,999.00	99,999.00
bicycle_shop_p_InArea	0.19	0.39	0	0	1
biergarten_p_InArea	0.05	0.22	0	0	1
bookshop_p_Distance_min	79,132.82	40,623.33	21.49	99,999.00	99,999.00
bookshop_p_InArea	0.21	0.41	0	0	1
bus_stop_p_Distance_min	15,464.93	36,071.72	3.72	85.18	99,999.00
bus_stop_p_InArea	0.85	0.36	0	1	1
butcher_p_Distance_min	86,377.71	34,294.79	23.78	99,999.00	99,999.00
butcher_p_InArea	0.14	0.34	0	0	1
cafe_p_Distance_min	43,901.03	49,591.92	4.13	242.44	99,999.00
cafe_p_InArea	0.56	0.50	0	1	1
car_dealership_p_InArea	0.09	0.28	0	0	1
car_rental_p_InArea	0.07	0.26	0	0	1
car_sharing_p_Distance_min	84,928.84	35,768.01	4.12	99,999.00	99,999.00
car_sharing_p_InArea	0.15	0.36	0	0	1
car_wash_p_InArea	0.03	0.17	0	0	1
cemetery_a_Density	0.01	0.03	0.00	0.00	0.32
cemetery_a_Distance_min	88,991.09	31,285.57	21.80	99,999.00	99,999.00
chemist_p_Distance_min	79,853.50	40,102.15	12.74	99,999.00	99,999.00
chemist_p_InArea	0.20	0.40	0	0	1
cinema_p_InArea	0.06	0.23	0	0	1
clothes_p_Distance_min	71,154.35	45,295.23	11.55	99,999.00	99,999.00
clothes_p_InArea	0.29	0.45	0	0	1

Continued on next page

Table A.1 – Continued from previous page

Statistic	Mean	St. Dev.	Min	Median	Max
commercial_a_Density	0.05	0.12	0.00	0.00	0.76
commercial_a_Distance_min	67,972.61	46,639.14	0.16	99,999.00	99,999.00
community_centre_p_Distance_min	90,292.55	29,596.06	12.73	99,999.00	99,999.00
community_centre_p_InArea	0.10	0.30	0	0	1
computer_shop_p_Distance_min	91,596.26	27,735.59	20.24	99,999.00	99,999.00
computer_shop_p_InArea	0.08	0.28	0	0	1
convenience_p_Distance_min	74,788.18	43,404.18	6.22	99,999.00	99,999.00
convenience_p_InArea	0.25	0.43	0	0	1
crossing_p_Distance_min	8,058.17	27,099.57	1.84	58.85	99,999.00
cycleway_l_Density	1,641.67	2,292.14	0.00	739.36	14,661.32
cycleway_l_Distance_min	28,967.86	45,299.51	0.12	173.49	99,999.00
dentist_p_Distance_min	84,353.48	36,314.65	16.11	99,999.00	99,999.00
dentist_p_InArea	0.16	0.36	0	0	1
department_store_p_InArea	0.04	0.19	0	0	1
doctors_p_Distance_min	75,942.14	42,731.82	10.35	99,999.00	99,999.00
doctors_p_InArea	0.24	0.43	0	0	1
doityourself_p_InArea	0.05	0.22	0	0	1
drinking_water_p_InArea	0.04	0.19	0	0	1
farm_a_Density	0.01	0.04	0.00	0.00	0.50
fast_food_p_Distance_min	47,815.52	49,922.41	7.30	324.04	99,999.00
fast_food_p_InArea	0.52	0.50	0	1	1
fire_station_p_InArea	0.03	0.17	0	0	1
florist_p_Distance_min	76,522.09	42,374.48	7.25	99,999.00	99,999.00
florist_p_InArea	0.24	0.42	0	0	1
footway_l_Density	9,031.57	6,379.27	0.00	7,811.88	41,758.69
footway_l_Distance_min	318.91	5,382.37	0.01	12.44	99,999.00
forest_a_Density	0.02	0.07	0.00	0.00	0.68
forest_a_Distance_min	72,324.43	44,719.26	0.65	99,999.00	99,999.00
fountain_p_Distance_min	79,711.35	40,204.03	6.85	99,999.00	99,999.00
fountain_p_InArea	0.20	0.40	0	0	1
fuel_a_Density	0.0002	0.001	0.00	0.00	0.01
fuel_a_Distance_min	90,295.58	29,586.80	13.90	99,999.00	99,999.00
fuel_p_Distance_min	87,110.65	33,488.32	27.22	99,999.00	99,999.00
fuel_p_InArea	0.13	0.34	0	0	1
furniture_shop_p_Distance_min	85,802.01	34,889.34	19.41	99,999.00	99,999.00
furniture_shop_p_InArea	0.14	0.35	0	0	1
grass_a_Density	0.01	0.03	0.00	0.00	0.31
grass_a_Distance_min	52,312.04	49,919.34	0.09	99,999.00	99,999.00
guesthouse_p_InArea	0.02	0.16	0	0	1
hairdresser_p_Distance_min	51,450.28	49,942.79	4.30	99,999.00	99,999.00
hairdresser_p_InArea	0.49	0.50	0	0	1
hotel_p_Distance_min	70,731.72	45,476.49	10.20	99,999.00	99,999.00
hotel_p_InArea	0.29	0.46	0	0	1
industrial_a_Density	0.04	0.13	0.00	0.00	1
industrial_a_Distance_min	80,153.50	39,864.56	0.09	99,999.00	99,999.00
jeweller_p_Distance_min	84,493.47	36,188.61	11.14	99,999.00	99,999.00
jeweller_p_InArea	0.16	0.36	0	0	1
kindergarten_p_Distance_min	59,728.61	49,000.44	1.27	99,999.00	99,999.00
kindergarten_p_InArea	0.40	0.49	0	0	1
kiosk_p_Distance_min	57,248.60	49,441.66	7.59	99,999.00	99,999.00
kiosk_p_InArea	0.43	0.50	0	0	1
laundry_p_Distance_min	79,279.47	40,515.74	9.36	99,999.00	99,999.00
laundry_p_InArea	0.21	0.41	0	0	1
library_p_Distance_min	82,899.01	37,643.37	8.94	99,999.00	99,999.00
library_p_InArea	0.17	0.38	0	0	1
living_street_l_Density	259.95	709.54	0.00	0.00	7,097.32
living_street_l_Distance_min	72,469.38	44,646.71	1.90	99,999.00	99,999.00
meadow_a_Density	0.003	0.02	0.00	0.00	0.25
meadow_a_Distance_min	86,674.37	33,967.53	8.02	99,999.00	99,999.00
memorial_p_Distance_min	45,934.25	49,798.82	2.24	272.25	99,999.00
mobile_phone_shop_p_Distance_min	87,536.13	33,025.01	17.37	99,999.00	99,999.00

Continued on next page

Table A.1 – Continued from previous page

Statistic	Mean	St. Dev.	Min	Median	Max
mobile_phone_shop_p_InArea	0.12	0.33	0	0	1
museum_p_Distance_min	91,450.71	27,953.68	24.50	99,999.00	99,999.00
museum_p_InArea	0.09	0.28	0	0	1
newsagent_p_InArea	0.07	0.25	0	0	1
nightclub_p_InArea	0.07	0.26	0	0	1
optician_p_Distance_min	81,881.94	38,511.27	10.63	99,999.00	99,999.00
optician_p_InArea	0.18	0.39	0	0	1
outdoor_shop_p_InArea	0.03	0.17	0	0	1
park_a_Density	0.06	0.11	0.00	0.02	0.99
park_a_Distance_min	25,052.13	43,260.21	0.13	154.42	99,999.00
parking_a_Density	0.01	0.02	0.00	0.01	0.21
parking_a_Distance_min	20,109.54	40,010.08	0.05	109.84	99,999.00
parking_bicycle_a_Density	0.0001	0.001	0.00	0.00	0.02
parking_bicycle_a_Distance_min	84,634.18	36,059.90	2.45	99,999.00	99,999.00
parking_bicycle_p_Distance_min	41,417.47	49,238.89	0.75	196.08	99,999.00
parking_bicycle_p_InArea	0.59	0.49	0	1	1
parking_multistorey_a_Density	0.002	0.01	0.00	0.00	0.08
parking_multistorey_a_Distance_min	81,163.05	39,087.45	1.05	99,999.00	99,999.00
parking_multistorey_p_InArea	0.07	0.25	0	0	1
parking_p_Distance_min	62,182.90	48,461.19	3.07	99,999.00	99,999.00
parking_p_InArea	0.38	0.49	0	0	1
parking_underground_p_Distance_min	73,195.39	44,272.94	1.31	99,999.00	99,999.00
parking_underground_p_InArea	0.27	0.44	0	0	1
path_l_Density	912.74	1,560.32	0.00	262.38	11,925.40
path_l_Distance_min	34,347.85	47,420.41	0.65	210.25	99,999.00
pedestrian_l_Density	2,208.13	5,608.69	0.00	240.69	61,134.22
pedestrian_l_Distance_min	45,041.40	49,739.90	0.002	271.37	99,999.00
pharmacy_p_Distance_min	55,358.47	49,685.79	5.20	99,999.00	99,999.00
pharmacy_p_InArea	0.45	0.50	0	0	1
picnic_site_p_InArea	0.02	0.15	0	0	1
pitch_p_InArea	0.06	0.24	0	0	1
playground_p_Distance_min	73,785.47	43,947.46	25.01	99,999.00	99,999.00
playground_p_InArea	0.26	0.44	0	0	1
police_p_Distance_min	93,334.50	24,935.03	34.87	99,999.00	99,999.00
police_p_InArea	0.07	0.25	0	0	1
post_office_p_Distance_min	80,004.33	39,982.15	9.45	99,999.00	99,999.00
post_office_p_InArea	0.20	0.40	0	0	1
pub_p_Distance_min	61,164.67	48,708.55	10.28	99,999.00	99,999.00
pub_p_InArea	0.39	0.49	0	0	1
public_building_p_InArea	0.05	0.21	0	0	1
rail_l_Density	3,402.45	9,267.76	0.00	0.00	111,770.10
rail_l_Distance_min	66,081.15	47,330.59	0.02	99,999.00	99,999.00
railway_station_p_InArea	0.15	0.35	0	0	1
recreation_ground_a_Density	0.004	0.03	0.00	0.00	0.70
recycling_clothes_p_InArea	0.15	0.36	0	0	1
recycling_p_Distance_min	79,135.14	40,618.82	14.38	99,999.00	99,999.00
recycling_p_InArea	0.21	0.41	0	0	1
residential_a_Density	0.47	0.34	0.00	0.47	1
residential_a_Distance_min	35,763.15	47,902.60	0.07	147.36	99,999.00
residential_l_Density	7,299.02	4,536.12	0.00	7,080.67	22,512.81
residential_l_Distance_min	4,543.72	20,734.09	0.001	22.47	99,999.00
restaurant_p_Distance_min	21,862.73	41,249.64	7.25	123.41	99,999.00
restaurant_p_InArea	0.78	0.41	0	1	1
school_p_Distance_min	85,370.48	35,321.91	46.57	99,999.00	99,999.00
school_p_InArea	0.15	0.35	0	0	1
scrub_a_Density	0.01	0.03	0.00	0.00	0.40
scrub_a_Distance_min	59,282.59	49,097.83	0.22	99,999.00	99,999.00
secondary_l_Density	2,038.91	2,661.68	0.00	1,359.50	21,809.98
secondary_l_Distance_min	38,525.19	48,634.69	0.04	203.07	99,999.00
secondary_link_l_Density	37.60	143.02	0.00	0.00	1,878.98
secondary_link_l_Distance_min	86,815.54	33,820.10	2.95	99,999.00	99,999.00

Continued on next page

Table A.1 – Continued from previous page

Statistic	Mean	St. Dev.	Min	Median	Max
service_l.Density	4,329.81	3,408.17	0.00	3,595.22	21,876.76
service_l.Distance_min	3,537.12	18,338.67	0.06	40.40	99,999.00
shelter_p.InArea	0.10	0.31	0	0	1
shoe_shop_p.Distance_min	84,924.26	35,778.87	9.19	99,999.00	99,999.00
shoe_shop_p.InArea	0.15	0.36	0	0	1
sports_centre_p.Distance_min	85,802.03	34,889.28	10.82	99,999.00	99,999.00
sports_centre_p.InArea	0.14	0.35	0	0	1
sports_shop_p.InArea	0.08	0.28	0	0	1
stationery_p.Distance_min	90,727.04	28,995.94	27.18	99,999.00	99,999.00
stationery_p.InArea	0.09	0.29	0	0	1
steps_l.Density	453.22	1,091.34	0.00	112.65	13,483.70
steps_l.Distance_min	23,157.43	42,118.52	0.01	121.05	99,999.00
stream_l.Density	215.08	594.05	0.00	0.00	3,455.67
stream_l.Distance_min	83,917.72	36,723.06	6.11	99,999.00	99,999.00
supermarket_p.Distance_min	53,916.70	49,810.77	15.10	99,999.00	99,999.00
supermarket_p.InArea	0.46	0.50	0	0	1
taxi_p.Distance_min	75,356.32	43,088.64	4.43	99,999.00	99,999.00
taxi_p.InArea	0.25	0.43	0	0	1
tertiary_l.Density	1,341.28	1,995.53	0.00	400.61	16,603.28
tertiary_l.Distance_min	46,214.81	49,830.73	0.40	322.37	99,999.00
theatre_p.Distance_min	89,134.48	31,108.54	23.92	99,999.00	99,999.00
theatre_p.InArea	0.11	0.31	0	0	1
tourist_info_p.Distance_min	77,532.80	41,730.16	5.38	99,999.00	99,999.00
tourist_info_p.InArea	0.22	0.42	0	0	1
toy_shop_p.InArea	0.07	0.26	0	0	1
traffic_signals_p.Distance_min	14,305.54	34,920.75	0.45	85.91	99,999.00
traffic_signals_p.InArea	0.86	0.35	0	1	1
travel_agent_p.Distance_min	84,492.84	36,190.07	13.10	99,999.00	99,999.00
travel_agent_p.InArea	0.16	0.36	0	0	1
tree_p.Density	171.39	274.22	0.00	32.93	1,698.63
turning_circle_p.Distance_min	64,656.24	47,764.07	4.54	99,999.00	99,999.00
turning_circle_p.InArea	0.35	0.48	0	0	1
university_p.InArea	0.04	0.20	0	0	1
water_a.Density	0.01	0.03	0.00	0.00	0.34
water_a.Distance_min	71,744.73	45,003.46	6.22	99,999.00	99,999.00
City_center.Distance_min_all	3,535.73	2,821.42	59.65	2,843.88	16,644.05
trackAll_l.Density	286.02	959.22	0.00	0.00	8,482.23
Population	940,729.50	590,312.60	73,836	732,688	1,787,604

Appendix B

Analysis between the rentals and thier most correlated variables

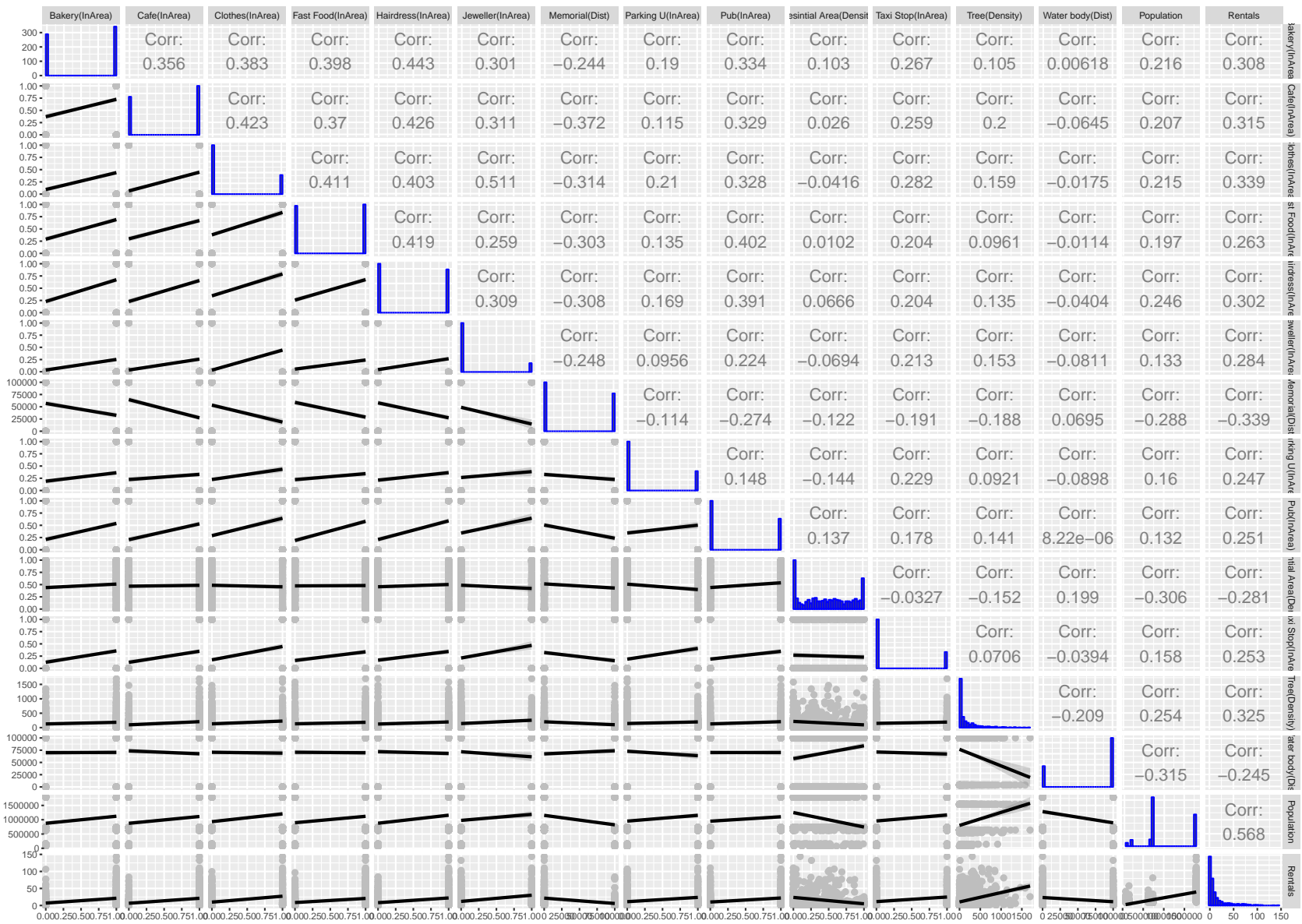


Figure B.1: Analysis of the most correlated variables

Appendix C

Analysis between the rentals and the most successful factors named on the literature

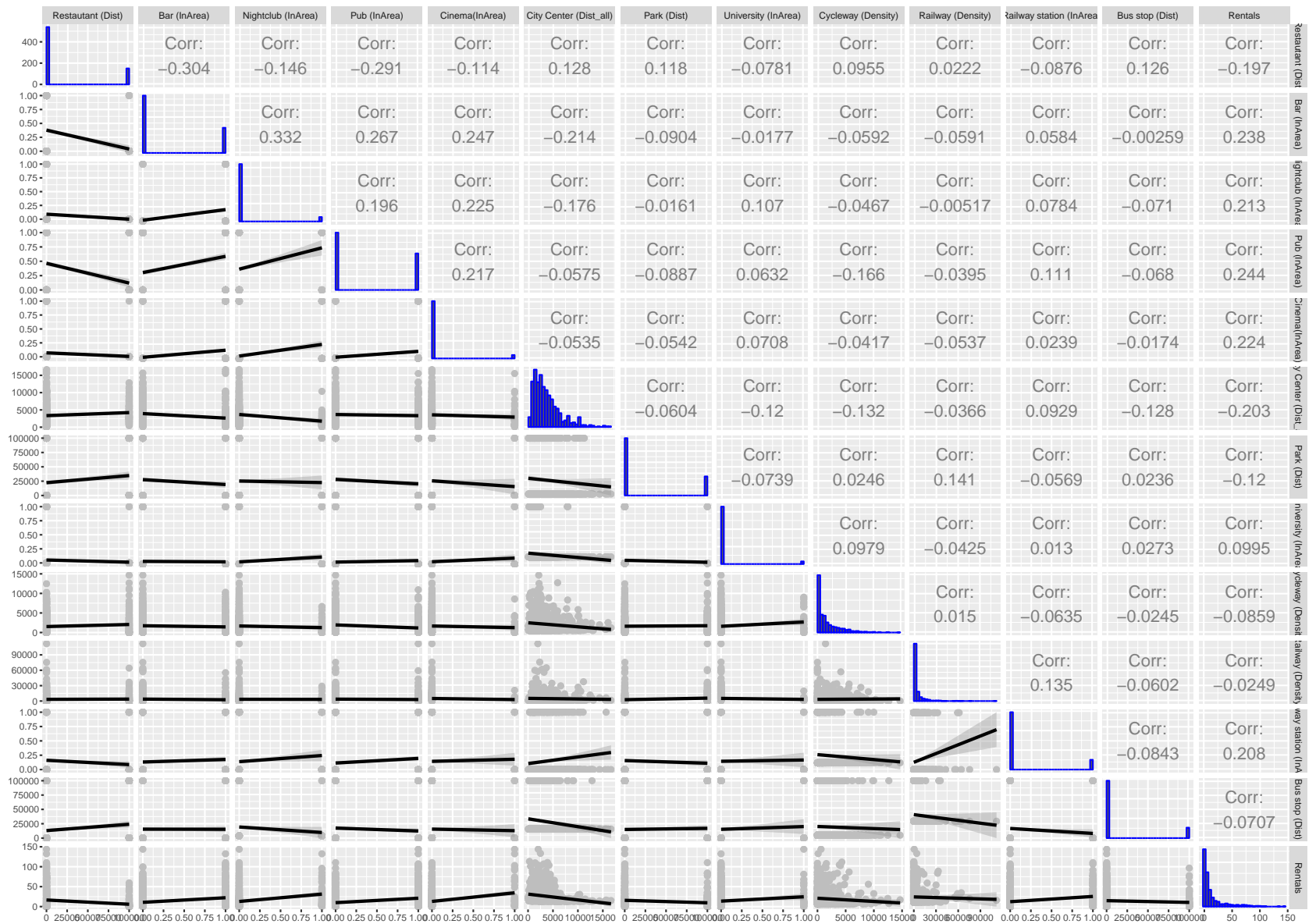


Figure C.1: Analysis between the rentals and the most successful factors named on the literature

Appendix D

Model assessment parameters

D.1 Stepwise regression

D.2 GLM + lasso

D.3 GBM

Table D.1: Results from stepwise regression (No transformations)

Time unit	# var.	MSE	R^2	R^2_{adj}	BIC	MSE (validation)	R^2 (validation)
WA1p	15	10.78	0.58	0.57	3378.26	7.58	0.27
WA2p	11	28.91	0.59	0.59	3984.68	17.18	0.23
WM1p	12	6.68	0.48	0.47	2996.37	1.53	0.23
WM2p	15	3.20	0.57	0.56	2591.08	1.45	0.21
WN1p	18	5.16	0.55	0.54	2912.12	3.38	0.33
WN2p	20	0.32	0.58	0.56	1117.62	0.38	0.25
SaA1p	14	22.64	0.53	0.52	3832.14	8.19	0.24
SaA2p	16	22.49	0.52	0.51	3846.74	9.10	0.16
SaM1p	16	0.24	0.51	0.49	929.46	0.12	0.24
SaM2p	17	3.51	0.53	0.52	2657.39	1.41	0.21
SaN1p	21	7.05	0.56	0.55	3111.92	3.57	0.33
SaN2p	12	9.48	0.47	0.46	3166.04	5.61	0.44
SuA1p	15	21.40	0.51	0.50	3803.09	14.82	0.07
SuA2p	16	21.04	0.49	0.48	3804.73	10.34	0.15
SuM1p	12	0.30	0.46	0.45	1020.06	0.17	0.10
SuM2p	14	2.51	0.52	0.51	2406.43	1.71	0.23
SuN1p	18	3.26	0.54	0.53	2609.44	2.01	0.25
SuN2p	17	9.55	0.50	0.49	3258.66	6.08	0.43
WA1a	13	10.41	0.57	0.56	3348.90	9.05	0.19
WA2a	16	33.13	0.56	0.55	4096.34	16.12	0.23
WM1a	11	7.16	0.45	0.44	2951.03	2.43	0.09
WM2a	18	4.34	0.58	0.56	2760.71	3.31	0.14
WN1a	18	5.71	0.56	0.55	2989.50	5.68	0.22
WN2a	18	0.30	0.56	0.55	1095.42	0.54	0.14
SaA1a	19	25.64	0.52	0.51	3948.26	9.19	0.13
SaA2a	16	24.63	0.53	0.52	3903.73	14.49	0.08
SaM1a	14	0.26	0.45	0.43	954.44	0.13	0.21
SaM2a	16	2.89	0.52	0.51	2509.96	1.46	0.12
SaN1a	19	9.18	0.55	0.54	3272.52	7.22	0.20
SaN2a	17	5.94	0.51	0.50	2998.40	4.09	0.29
SuA1a	13	24.59	0.46	0.45	3865.26	10.37	0.17
SuA2a	16	21.93	0.52	0.51	3836.81	14.57	0.11
SuM1a	8	0.38	0.37	0.36	1146.00	0.25	0.09
SuM2a	9	1.94	0.48	0.48	2200.41	1.30	0.10
SuN1a	17	3.43	0.56	0.55	2651.14	3.15	0.20
SuN2a	20	5.83	0.53	0.52	2987.99	5.05	0.21

Table D.2: Results from stepwise regression (Log transformation)

Time unit	# Var.	MSE	R^2	R^2_{adj}	BIC	MSE (validation)	R^2 (validation)
WA1p	19	1.1	0.7	0.70	1946	2.32	0.53
WA2p	16	1.0	0.7	0.71	1921	2.06	0.47
WM1p	16	1.2	0.7	0.65	1969	1.63	0.30
WM2p	13	1.0	0.7	0.69	1844	2.29	0.41
WN1p	15	1.0	0.7	0.68	1891	2.26	0.50
WN2p	19	0.9	0.7	0.68	1711	2.77	0.48
SaA1p	17	0.8	0.8	0.75	1792	3.10	0.45
SaA2p	16	0.9	0.7	0.74	1825	3.36	0.44
SaM1p	18	1.0	0.6	0.62	1809	2.10	0.34
SaM2p	16	1.0	0.7	0.69	1874	2.55	0.38
SaN1p	17	0.9	0.7	0.72	1813	2.70	0.54
SaN2p	13	1.1	0.7	0.69	1858	3.06	0.65
SuA1p	16	0.8	0.7	0.74	1783	2.86	0.29
SuA2p	19	0.9	0.7	0.73	1837	2.52	0.35
SuM1p	17	1.0	0.7	0.64	1728	1.89	0.30
SuM2p	18	1.0	0.7	0.67	1845	2.42	0.29
SuN1p	16	0.9	0.7	0.70	1794	2.07	0.45
SuN2p	14	1.0	0.7	0.74	1834	3.68	0.61
WA1a	18	1.1	0.7	0.72	1943	2.33	0.47
WA2a	19	0.9	0.7	0.71	1874	2.25	0.37
WM1a	20	1.4	0.7	0.65	2048	1.94	0.26
WM2a	14	1.2	0.7	0.69	1935	1.54	0.47
WN1a	15	1.0	0.7	0.65	1889	2.89	0.46
WN2a	19	1.0	0.6	0.63	1795	2.14	0.43
SaA1a	15	1.0	0.7	0.74	1854	2.74	0.41
SaA2a	15	0.9	0.7	0.74	1824	3.17	0.45
SaM1a	14	0.9	0.6	0.64	1709	1.99	0.37
SaM2a	17	0.9	0.7	0.73	1826	2.27	0.41
SaN1a	17	0.9	0.7	0.72	1815	2.62	0.47
SaN2a	19	1.1	0.7	0.67	1937	2.77	0.47
SuA1a	18	0.8	0.8	0.74	1790	2.81	0.32
SuA2a	22	0.9	0.8	0.74	1833	2.25	0.39
SuM1a	16	0.9	0.7	0.65	1715	3.30	0.14
SuM2a	14	0.9	0.7	0.70	1788	2.83	0.33
SuN1a	15	0.9	0.7	0.68	1820	2.53	0.34
SuN2a	14	1.0	0.7	0.67	1871	3.44	0.44

Table D.3: Results from stepwise regression (BoxCox transformation)

Time unit	# var.	MSE	R^2	R^2_{adj}	BIC	MSE (validation)	R^2 (validation)
WA1p	15	0.65	0.73	0.72	1616	1.60	0.50
WA2p	16	1.65	0.75	0.74	2213	4.10	0.43
WM1p	17	0.36	0.68	0.67	1232	0.61	0.30
WM2p	15	0.27	0.72	0.71	1047	0.56	0.41
WN1p	13	0.30	0.70	0.69	1107	1.08	0.53
WN2p	25	0.02	0.72	0.71	570	0.05	0.48
SaA1p	22	0.74	0.78	0.77	1735	2.41	0.45
SaA2p	14	0.71	0.76	0.75	1661	3.56	0.43
SaM1p	17	0.02	0.64	0.63	594	0.04	0.31
SaM2p	16	0.18	0.71	0.71	811	0.46	0.38
SaN1p	15	0.29	0.73	0.73	1095	1.21	0.52
SaN2p	17	0.18	0.72	0.71	787	0.79	0.66
SuA1p	21	0.68	0.77	0.76	1675	2.67	0.27
SuA2p	20	0.67	0.75	0.75	1669	2.31	0.34
SuM1p	17	0.01	0.67	0.66	850	0.02	0.30
SuM2p	19	0.13	0.70	0.69	617	0.30	0.31
SuN1p	21	0.15	0.73	0.72	731	0.34	0.44
SuN2p	20	0.15	0.76	0.75	703	1.02	0.66
WA1a	19	0.54	0.75	0.74	1521	1.24	0.45
WA2a	18	1.72	0.75	0.74	2250	3.16	0.40
WM1a	18	0.24	0.67	0.66	963	0.53	0.19
WM2a	16	0.23	0.72	0.71	954	0.32	0.41
WN1a	17	0.42	0.69	0.68	1345	1.15	0.46
WN2a	19	0.03	0.66	0.65	391	0.08	0.39
SaA1a	20	0.57	0.78	0.77	1564	1.67	0.36
SaA2a	16	0.75	0.76	0.76	1711	2.87	0.41
SaM1a	14	0.01	0.66	0.65	751	0.03	0.35
SaM2a	19	0.10	0.75	0.75	470	0.23	0.36
SaN1a	19	0.32	0.74	0.73	1185	0.93	0.48
SaN2a	18	0.26	0.69	0.68	1055	1.23	0.49
SuA1a	18	0.55	0.76	0.75	1521	1.73	0.32
SuA2a	19	0.79	0.76	0.75	1762	2.58	0.33
SuM1a	16	0.01	0.66	0.65	953	4.38	0.14
SuM2a	15	0.08	0.72	0.72	283	0.24	0.26
SuN1a	16	0.23	0.70	0.70	953	0.58	0.35
SuN2a	18	0.26	0.70	0.69	1034	1.35	0.51

Table D.4: Results from GLM (No transformations)

Time unit	λ	# var.	MSE	R^2	R^2_{adj}	BIC	MSE (validation)	R^2 (validation)
WA1p	0.59	12	13.63	0.46	0.45	1718	3.88	0.32
WA2p	0.81	13	38.09	0.46	0.45	2377	5.81	0.34
WM1p	0.45	10	8.56	0.33	0.32	1385	0.26	0.21
WM2p	0.27	20	4.00	0.46	0.44	991	0.65	0.32
WN1p	0.37	16	7.13	0.38	0.36	1327	1.65	0.43
WN2p	0.08	21	0.48	0.37	0.35	293	0.22	0.19
SaA1p	0.96	14	28.90	0.40	0.39	2199	3.07	0.24
SaA2p	1.00	10	30.43	0.35	0.34	2209	4.29	0.19
SaM1p	0.07	22	0.29	0.40	0.37	588	0.06	0.28
SaM2p	0.34	12	4.74	0.37	0.36	1043	0.55	0.27
SaN1p	0.46	26	8.90	0.45	0.42	1522	1.48	0.38
SaN2p	0.88	3	14.91	0.17	0.16	1657	2.32	0.35
SuA1p	1.03	9	29.17	0.34	0.33	2173	5.05	0.16
SuA2p	1.04	19	25.50	0.38	0.36	2156	4.88	0.19
SuM1p	0.14	5	0.44	0.19	0.18	443	0.05	0.07
SuM2p	0.26	24	2.82	0.46	0.43	791	0.52	0.28
SuN1p	0.35	23	4.41	0.38	0.35	1067	0.84	0.23
SuN2p	0.77	3	16.59	0.14	0.13	1752	2.63	0.25
WA1a	0.62	12	14.06	0.42	0.40	1740	5.63	0.22
WA2a	0.89	23	39.87	0.47	0.45	2467	9.25	0.21
WM1a	0.68	5	9.99	0.23	0.23	1408	0.62	0.03
WM2a	0.45	12	5.67	0.44	0.43	1141	2.07	0.15
WN1a	0.26	35	6.39	0.51	0.48	1386	3.11	0.32
WN2a	0.07	32	0.36	0.47	0.44	405	0.36	0.29
SaA1a	1.32	8	37.10	0.31	0.30	2321	3.44	0.10
SaA2a	0.97	23	31.12	0.41	0.39	2307	7.01	0.13
SaM1a	0.13	2	0.38	0.17	0.17	546	0.09	0.14
SaM2a	0.38	11	3.96	0.34	0.33	918	0.63	0.13
SaN1a	0.55	9	15.12	0.26	0.25	1747	3.24	0.09
SaN2a	0.33	25	6.99	0.43	0.40	1374	3.42	0.30
SuA1a	1.17	13	30.13	0.34	0.33	2212	5.05	0.17
SuA2a	0.78	23	26.54	0.42	0.40	2210	7.39	0.16
SuM1a	0.34	0	0.60	0.00	0.00	295	0.07	0.00
SuM2a	0.28	14	2.44	0.35	0.34	633	0.56	0.16
SuN1a	0.21	23	4.14	0.47	0.45	1033	1.73	0.24
SuN2a	0.34	15	8.51	0.32	0.30	1424	4.24	0.20

Table D.5: Results from GLM (Logarithmic transformation)

Time unit	λ	# var.	MSE	R^2	R^2_{adj}	BIC	MSE (validation)	R^2 (validation)
WA1p	0.12	31	1.18	0.68	0.66	304	1.97	0.51
WA2p	0.12	28	1.13	0.69	0.68	259	1.61	0.50
WM1p	0.12	27	1.38	0.61	0.59	369	1.44	0.26
WM2p	0.11	23	1.10	0.66	0.64	208	1.51	0.40
WN1p	0.12	28	1.13	0.66	0.64	258	1.94	0.58
WN2p	0.08	39	0.94	0.68	0.65	214	2.41	0.60
SaA1p	0.10	36	0.95	0.73	0.71	198	2.16	0.48
SaA2p	0.12	31	1.03	0.71	0.69	217	2.66	0.47
SaM1p	0.09	30	1.16	0.59	0.57	277	1.90	0.39
SaM2p	0.11	25	1.15	0.65	0.64	248	1.84	0.38
SaN1p	0.10	30	1.02	0.69	0.68	204	2.42	0.56
SaN2p	0.11	32	1.13	0.68	0.66	280	2.29	0.69
SuA1p	0.12	23	0.96	0.71	0.70	125	2.57	0.33
SuA2p	0.12	26	1.04	0.70	0.68	189	2.28	0.42
SuM1p	0.08	42	1.02	0.63	0.60	278	2.07	0.30
SuM2p	0.13	21	1.17	0.61	0.60	230	2.20	0.30
SuN1p	0.11	31	0.98	0.68	0.66	188	1.89	0.52
SuN2p	0.13	27	1.06	0.72	0.71	208	2.92	0.65
WA1a	0.10	38	1.12	0.71	0.69	316	1.95	0.43
WA2a	0.12	31	1.06	0.69	0.67	237	1.69	0.41
WM1a	0.10	35	1.57	0.63	0.61	492	1.30	0.37
WM2a	0.11	29	1.25	0.68	0.66	321	1.47	0.44
WN1a	0.11	28	1.11	0.63	0.61	244	2.07	0.51
WN2a	0.10	33	1.07	0.60	0.58	250	2.52	0.40
SaA1a	0.13	28	1.07	0.72	0.70	224	2.00	0.44
SaA2a	0.13	28	1.03	0.71	0.70	199	2.38	0.45
SaM1a	0.10	34	1.01	0.62	0.60	219	1.84	0.43
SaM2a	0.11	35	1.02	0.71	0.70	235	1.74	0.46
SaN1a	0.10	31	0.98	0.70	0.69	187	2.12	0.54
SaN2a	0.10	35	1.17	0.64	0.62	322	2.46	0.53
SuA1a	0.11	32	0.95	0.72	0.71	171	2.25	0.35
SuA2a	0.11	25	1.04	0.70	0.68	186	2.33	0.38
SuM1a	0.11	23	1.06	0.61	0.59	179	3.07	0.15
SuM2a	0.13	26	1.07	0.67	0.65	209	2.16	0.33
SuN1a	0.09	42	0.97	0.67	0.65	249	1.92	0.42
SuN2a	0.10	34	1.09	0.66	0.64	271	2.63	0.52

Table D.6: Results from GLM (Boxcox transformation)

Time unit	λ	# var.	MSE	R^2	R^2_{adj}	BIC	MSE (validation)	R^2 (validation)
WA1p	0.09	32	0.69	0.70	0.69	22.806	1.51	0.51
WA2p	0.15	31	1.83	0.72	0.71	579.957	3.00	0.51
WM1p	0.05	30	0.40	0.64	0.63	375.466	0.49	0.30
WM2p	0.06	24	0.30	0.67	0.66	585.854	0.39	0.40
WN1p	0.06	32	0.32	0.68	0.66	502.91	0.80	0.58
WN2p	0.01	50	0.02	0.71	0.68	2023.779	0.05	0.58
SaA1p	0.11	32	0.87	0.74	0.72	118.24	2.13	0.48
SaA2p	0.10	34	0.77	0.73	0.72	53.097	2.82	0.46
SaM1p	0.01	31	0.02	0.61	0.59	2125.465	0.03	0.41
SaM2p	0.04	30	0.20	0.68	0.67	806.771	0.34	0.38
SaN1p	0.05	37	0.30	0.72	0.70	499.95	0.97	0.55
SaN2p	0.05	31	0.19	0.69	0.68	801.48	0.77	0.70
SuA1p	0.10	25	0.79	0.73	0.72	10.1	2.65	0.33
SuA2p	0.11	27	0.79	0.71	0.70	25.762	2.18	0.40
SuM1p	0.01	45	0.01	0.65	0.63	2286.323	0.02	0.31
SuM2p	0.04	25	0.15	0.65	0.63	989.084	0.30	0.31
SuN1p	0.05	30	0.18	0.69	0.67	878.289	0.37	0.50
SuN2p	0.05	29	0.17	0.73	0.72	920.151	1.07	0.66
WA1a	0.07	39	0.57	0.73	0.71	101.183	1.07	0.43
WA2a	0.14	32	1.94	0.71	0.69	622.481	2.46	0.40
WM1a	0.04	35	0.26	0.65	0.63	587.013	0.26	0.38
WM2a	0.05	27	0.25	0.69	0.68	668.516	0.23	0.44
WN1a	0.06	32	0.45	0.65	0.64	289.735	0.95	0.50
WN2a	0.02	31	0.03	0.63	0.61	1954.314	0.08	0.39
SaA1a	0.10	32	0.67	0.74	0.72	48.201	1.28	0.43
SaA2a	0.09	38	0.81	0.74	0.72	113.867	2.43	0.43
SaM1a	0.01	30	0.02	0.63	0.61	2257.844	0.03	0.42
SaM2a	0.04	34	0.11	0.73	0.71	1130.723	0.20	0.44
SaN1a	0.06	32	0.35	0.72	0.70	448.878	0.89	0.51
SaN2a	0.04	42	0.28	0.67	0.64	527.113	1.09	0.52
SuA1a	0.08	31	0.62	0.73	0.72	100.641	1.52	0.34
SuA2a	0.10	32	0.89	0.73	0.71	134.842	2.40	0.37
SuM1a	0.10	26	1.04	0.61	0.60	188.057	3.06	0.15
SuM2a	0.04	26	0.09	0.68	0.66	1291.831	0.18	0.33
SuN1a	0.03	49	0.23	0.70	0.67	593.375	0.47	0.41
SuN2a	0.05	36	0.28	0.67	0.65	562.942	1.22	0.50

Table D.7: Results from GBM (No transformations)

Time unit	N. of trees	MSE	R^2	R^2_{adj}	BIC	MSE (validation)	R^2 (validation)
WA1p	5148	1.9	0.86	0.81	1735	3.66	0.29
WA2p	9120	2.01	0.94	0.93	1809	5.04	0.35
WM1p	6342	1.35	0.86	0.81	1297	0.74	0.33
WM2p	8750	0.71	0.93	0.91	506	0.65	0.33
WN1p	9995	0.89	0.93	0.91	776	1.24	0.47
WN2p	7930	0.29	0.89	0.86	530	0.22	0.21
SaA1p	6348	2.16	0.9	0.87	1893	2.43	0.22
SaA2p	4992	2.59	0.86	0.81	2124	3.29	0.2
SaM1p	5624	0.24	0.88	0.84	736	0.05	0.36
SaM2p	5663	0.83	0.91	0.88	699	0.45	0.27
SaN1p	9975	1.02	0.93	0.92	955	1.07	0.44
SaN2p	9820	1.61	0.85	0.81	1501	3.56	0.4
SuA1p	9589	1.76	0.93	0.91	1639	4.11	0.12
SuA2p	9986	1.75	0.93	0.9	1633	3.93	0.16
SuM1p	8207	0.3	0.83	0.78	477	0.05	0.11
SuM2p	8492	0.57	0.94	0.92	236	0.41	0.25
SuN1p	9960	0.71	0.93	0.91	499	0.82	0.28
SuN2p	9786	1.57	0.87	0.83	1483	4.18	0.33
WA1a	4839	1.83	0.86	0.82	1689	5.61	0.18
WA2a	9580	2.05	0.94	0.93	1831	9.31	0.16
WM1a	3072	2.1	0.66	0.56	1805	1.05	0.18
WM2a	4486	1.34	0.83	0.77	1279	2.18	0.14
WN1a	6961	1.01	0.92	0.9	939	3.24	0.25
WN2a	5643	0.27	0.89	0.86	640	0.39	0.22
SaA1a	9999	2.26	0.9	0.88	1949	5.52	0.11
SaA2a	9995	1.82	0.94	0.92	1682	5.8	0.13
SaM1a	5856	0.29	0.82	0.75	517	0.07	0.24
SaM2a	7754	0.75	0.91	0.88	563	0.75	0.13
SaN1a	9920	1.32	0.92	0.89	1268	2.56	0.3
SaN2a	9998	0.96	0.92	0.9	877	3.05	0.35
SuA1a	3384	3.25	0.77	0.7	2402	4.15	0.08
SuA2a	9730	1.73	0.93	0.91	1620	6.45	0.12
SuM1a	3490	0.5	0.59	0.45	105	0.06	0.04
SuM2a	5897	0.66	0.88	0.85	416	0.47	0.13
SuN1a	6811	0.81	0.91	0.89	671	1.58	0.25
SuN2a	7691	1.09	0.91	0.88	1030	3.39	0.31

Table D.8: Results from GBM (Logarithmic transformation)

Time unit	N. of trees	MSE	R^2	R^2_{adj}	BIC	MSE (validation)	R^2 (validation)
WA1p	6380	0.63	0.89	0.86	355.74	0.95	0.51
WA2p	5998	0.64	0.89	0.86	364.35	0.83	0.46
WM1p	5530	0.73	0.85	0.8	532.34	1.22	0.43
WM2p	5813	0.64	0.87	0.83	370.88	0.85	0.47
WN1p	5308	0.67	0.86	0.82	430.49	0.91	0.56
WN2p	4967	0.64	0.86	0.81	404.39	1.14	0.64
SaA1p	6239	0.58	0.9	0.87	254.01	0.93	0.48
SaA2p	5308	0.64	0.88	0.85	364.22	1.15	0.46
SaM1p	5082	0.7	0.83	0.77	499.17	0.99	0.48
SaM2p	5745	0.65	0.87	0.83	400.23	0.78	0.43
SaN1p	7728	0.54	0.91	0.89	164.56	1.01	0.61
SaN2p	4797	0.7	0.86	0.82	494.09	1.03	0.72
SuA1p	7661	0.53	0.91	0.89	133.74	1.21	0.35
SuA2p	5574	0.62	0.89	0.85	326.6	1.01	0.44
SuM1p	4379	0.7	0.82	0.76	502.66	1.06	0.36
SuM2p	4922	0.68	0.85	0.8	441.5	0.94	0.40
SuN1p	7891	0.53	0.91	0.88	145.91	0.91	0.54
SuN2p	5639	0.63	0.89	0.86	356.46	1.38	0.70
WA1a	5665	0.67	0.88	0.85	417.34	1.13	0.45
WA2a	6459	0.61	0.89	0.86	306.1	1.15	0.40
WM1a	6908	0.74	0.87	0.83	555.31	1.37	0.36
WM2a	6113	0.68	0.88	0.84	446.95	1.39	0.43
WN1a	5936	0.65	0.86	0.82	386.79	1.24	0.56
WN2a	5552	0.63	0.85	0.8	374.95	1.34	0.46
SaA1a	6448	0.62	0.9	0.87	317.79	1.14	0.38
SaA2a	7144	0.57	0.91	0.88	219.87	1.24	0.41
SaM1a	4929	0.64	0.84	0.79	399.9	1.23	0.45
SaM2a	7890	0.54	0.92	0.89	174.28	1.09	0.40
SaN1a	5353	0.64	0.88	0.84	362.35	0.99	0.60
SaN2a	5917	0.67	0.86	0.82	420.68	1.17	0.61
SuA1a	9836	0.48	0.93	0.91	2.91	1.23	0.33
SuA2a	9488	0.49	0.93	0.91	34.37	1.33	0.34
SuM1a	4994	0.64	0.85	0.79	405.53	1.62	0.21
SuM2a	4489	0.68	0.86	0.81	444.12	1.32	0.31
SuN1a	5899	0.62	0.87	0.83	339.57	1.19	0.44
SuN2a	6591	0.61	0.88	0.85	314.48	1.31	0.60

Table D.9: Results from GBM (Boxcox transformation)

Time unit	N. of trees	MSE	R^2	R^2_{adj}	BIC	MSE (validation)	R^2 (validation)
WA1p	5337	0.53	0.88	0.84	139.78	0.79	0.53
WA2p	6162	0.82	0.9	0.87	684.97	1.56	0.48
WM1p	6736	0.37	0.88	0.84	299.29	0.21	0.44
WM2p	4848	0.36	0.86	0.82	334.6	0.19	0.48
WN1p	6070	0.34	0.88	0.85	411.74	0.48	0.58
WN2p	6637	0.08	0.89	0.86	2016.73	0.04	0.60
SaA1p	8014	0.5	0.93	0.9	48.73	0.99	0.47
SaA2p	5934	0.54	0.9	0.87	142.34	1.39	0.46
SaM1p	7261	0.08	0.88	0.84	2090.84	0.02	0.48
SaM2p	7902	0.24	0.91	0.88	840.77	0.15	0.43
SaN1p	5912	0.34	0.9	0.86	422.91	0.5	0.61
SaN2p	6470	0.25	0.9	0.86	736.77	0.52	0.72
SuA1p	6969	0.5	0.91	0.89	57.45	1.35	0.35
SuA2p	6503	0.51	0.91	0.88	79.63	1.09	0.42
SuM1p	4830	0.07	0.84	0.79	2125	0.01	0.36
SuM2p	4429	0.26	0.85	0.8	751.54	0.14	0.41
SuN1p	6311	0.25	0.89	0.86	805.03	0.21	0.52
SuN2p	6423	0.24	0.91	0.88	855.34	0.77	0.70
WA1a	6097	0.47	0.9	0.87	20.54	0.67	0.43
WA2a	6086	0.85	0.89	0.86	725.13	1.58	0.40
WM1a	5804	0.33	0.85	0.81	416.97	0.2	0.36
WM2a	5640	0.32	0.88	0.84	482.1	0.19	0.43
WN1a	5439	0.43	0.86	0.82	128.19	0.58	0.54
WN2a	4695	0.11	0.84	0.79	1740.37	0.07	0.45
SaA1a	5496	0.52	0.89	0.86	102.32	0.72	0.38
SaA2a	5946	0.56	0.9	0.87	197.66	1.33	0.40
SaM1a	4821	0.08	0.85	0.8	2050	0.02	0.45
SaM2a	6446	0.2	0.9	0.87	1059.4	0.12	0.41
SaN1a	5319	0.38	0.88	0.85	279.08	0.5	0.57
SaN2a	7051	0.3	0.89	0.86	555.86	0.74	0.60
SuA1a	5453	0.5	0.89	0.86	58.34	0.82	0.31
SuA2a	5915	0.58	0.9	0.87	242.03	1.37	0.33
SuM1a	3976	0.07	0.82	0.76	2171.4	0.02	0.20
SuM2a	6487	0.17	0.9	0.87	1239	0.11	0.31
SuN1a	5568	0.32	0.87	0.83	500.98	0.29	0.45
SuN2a	5043	0.34	0.86	0.82	411.87	0.87	0.59

Appendix E

Fitted vs observed model values

)

E.1 Stepwise regression

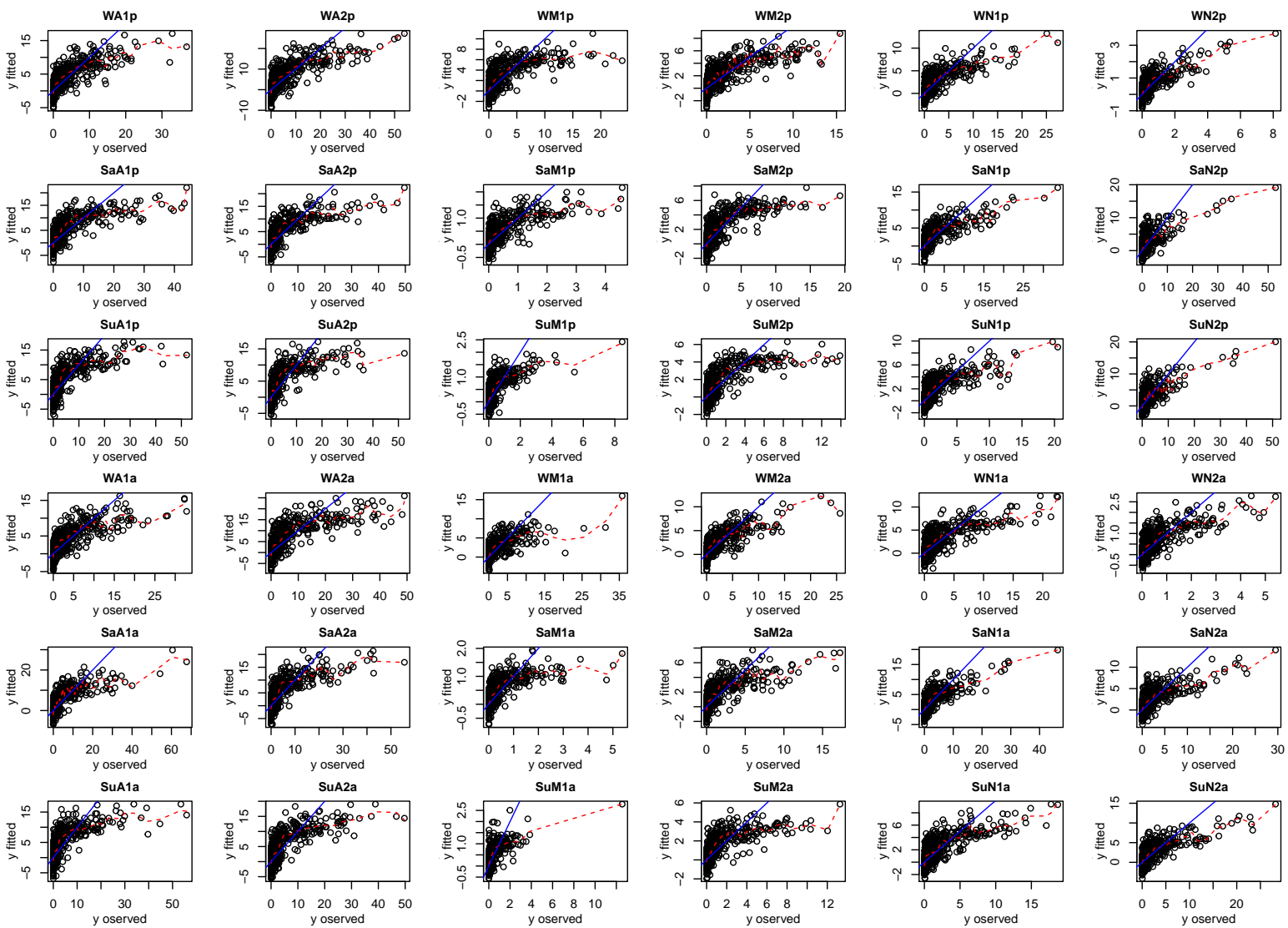


Figure E.1: Stepwise regression fitted vs observed values (No Treatment)

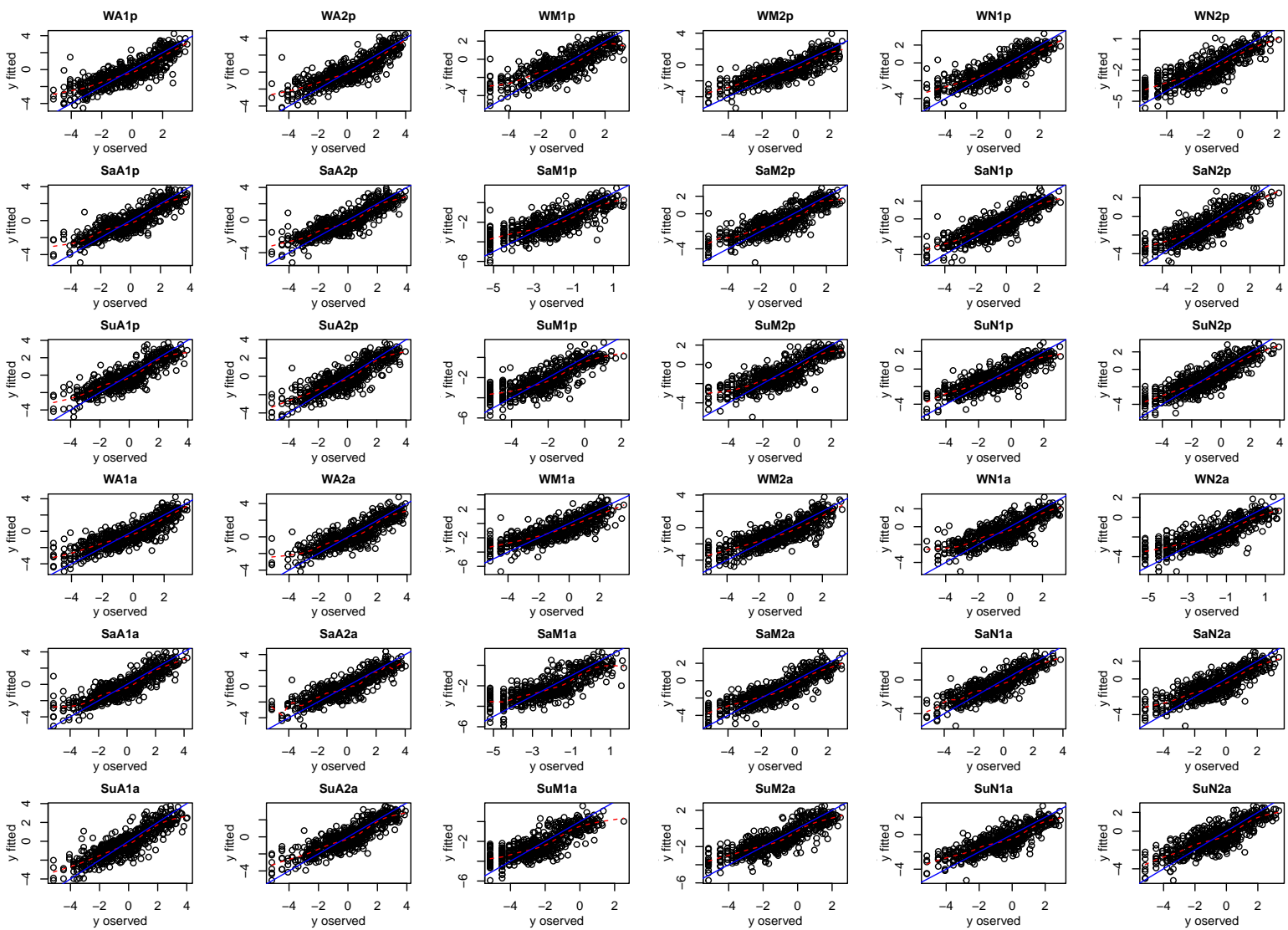


Figure E.2: Stepwise regression fitted vs observed values (Logarithmic transformation)

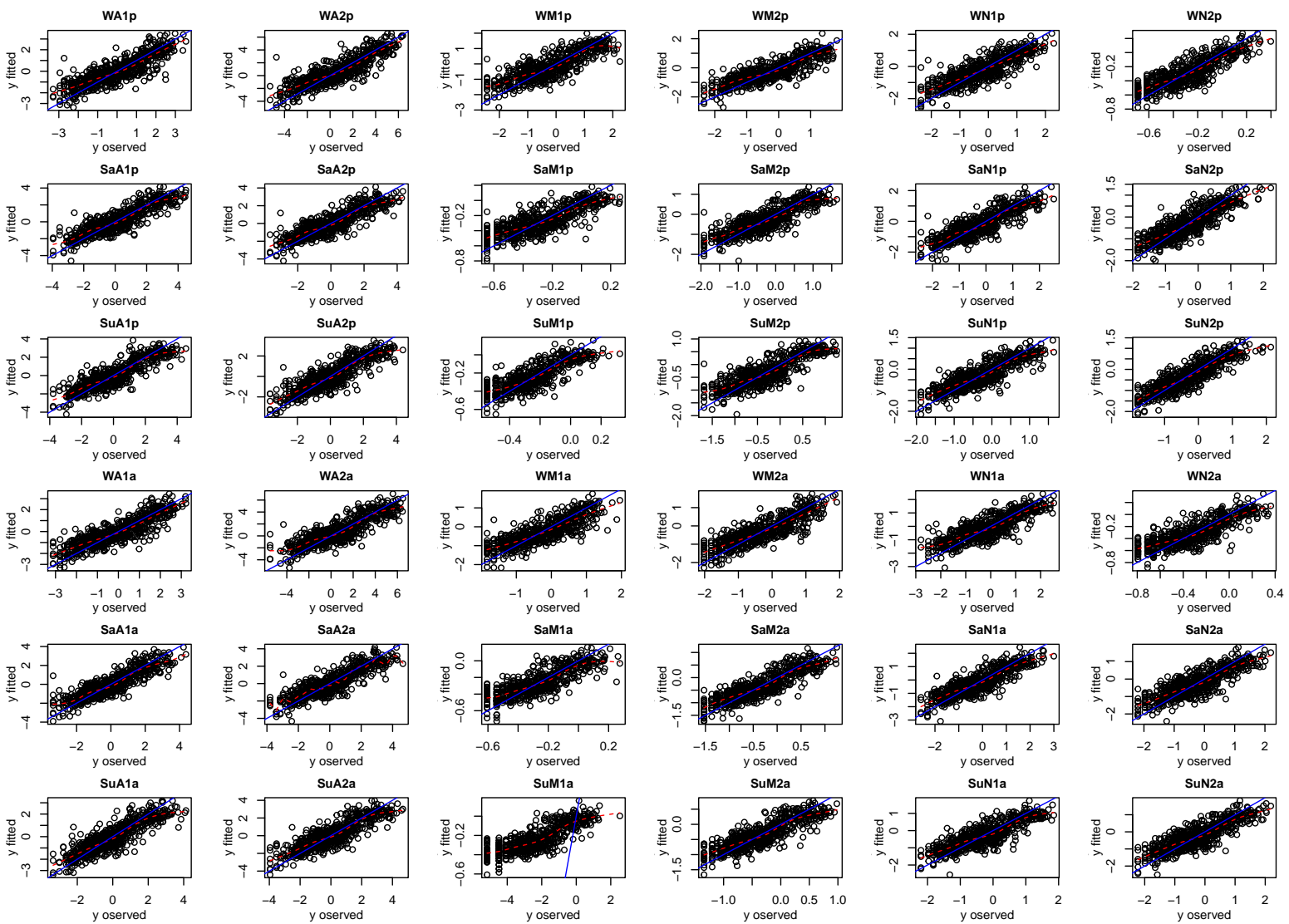


Figure E.3: Stepwise regression fitted vs observed values (Boxcox transformation)

E.2 GLM

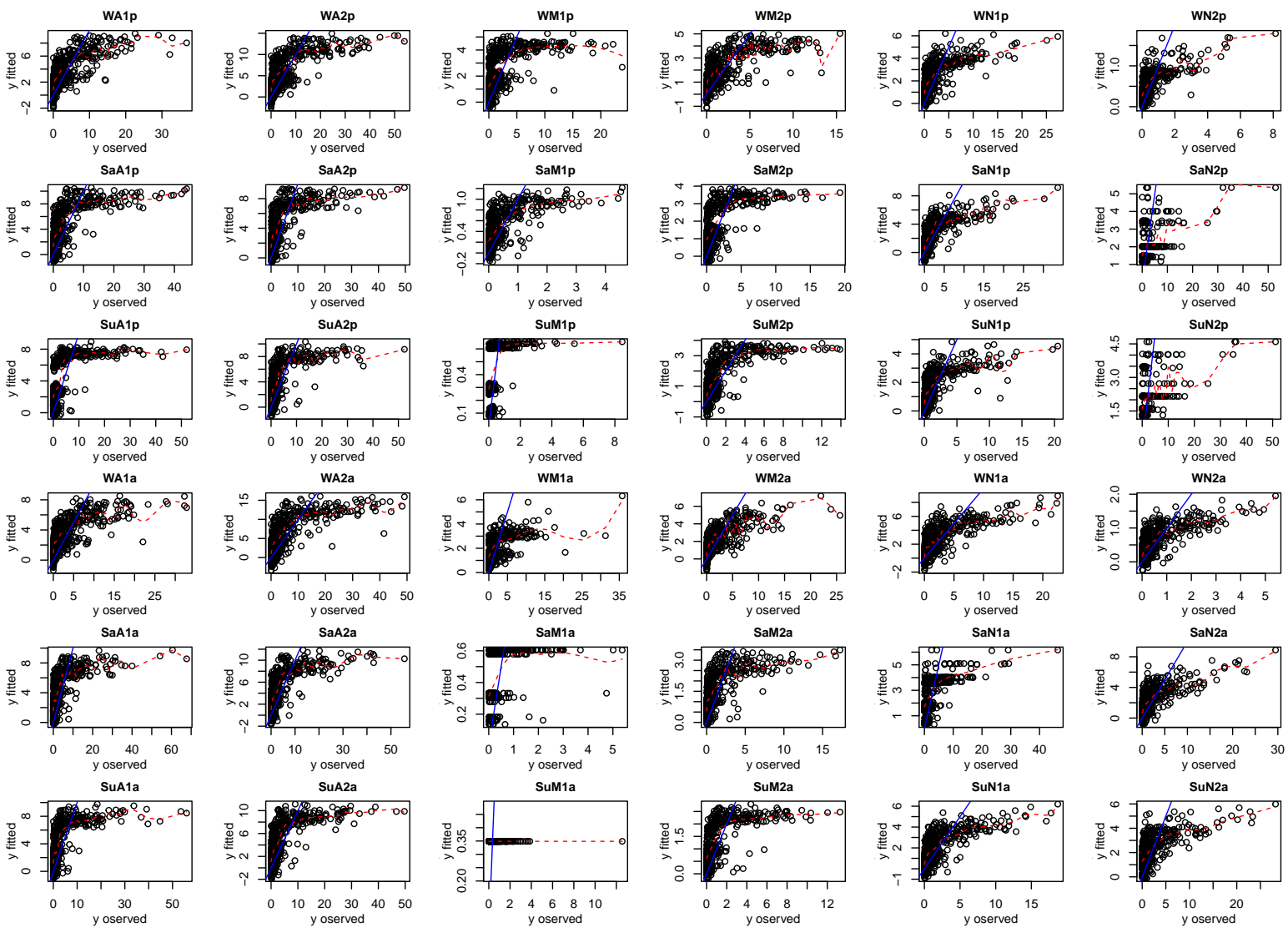


Figure E.4: GLM fitted vs observed values (No transformations)

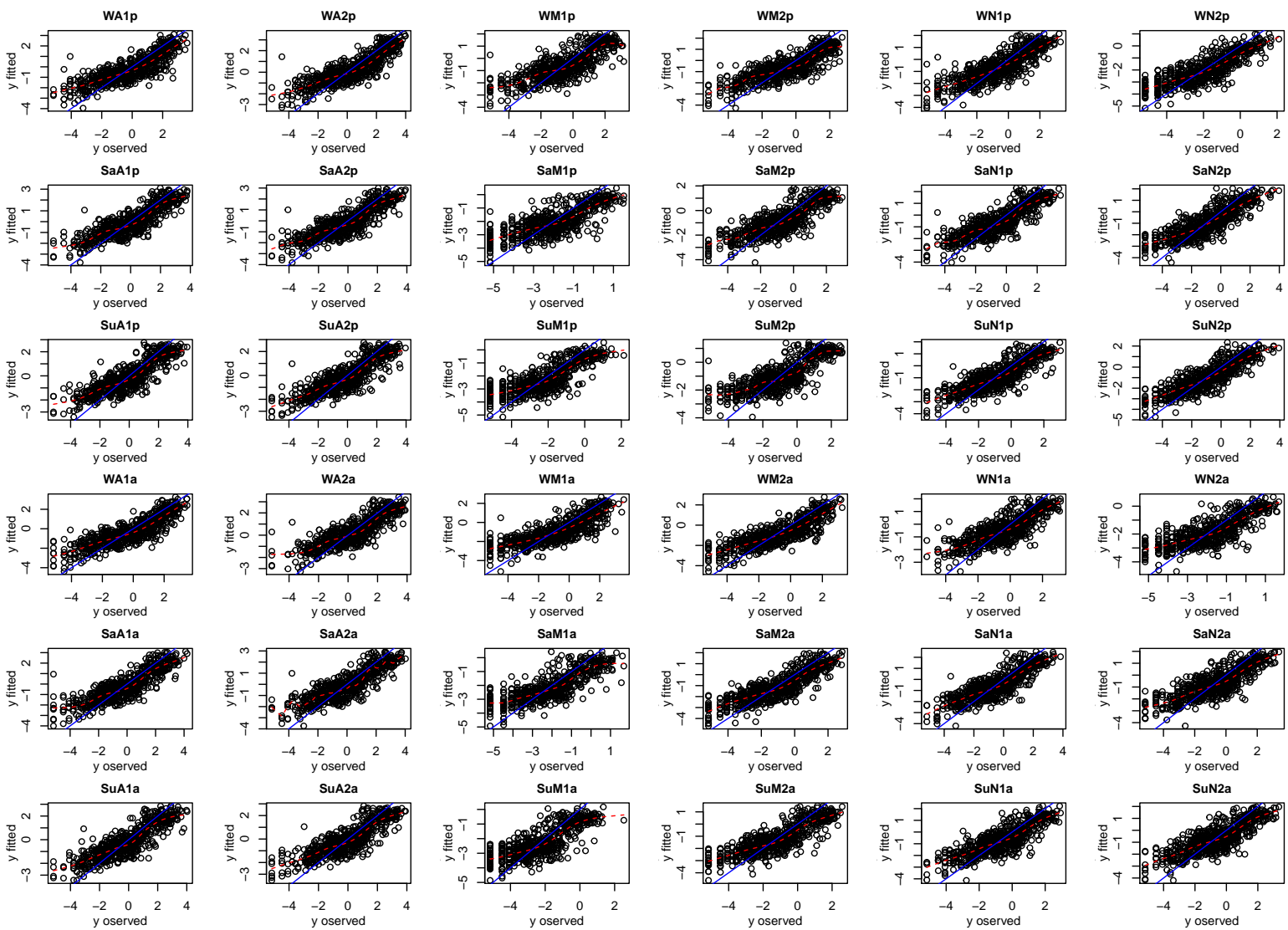


Figure E.5: GLM fitted vs observed values (Logarithmic transformation)

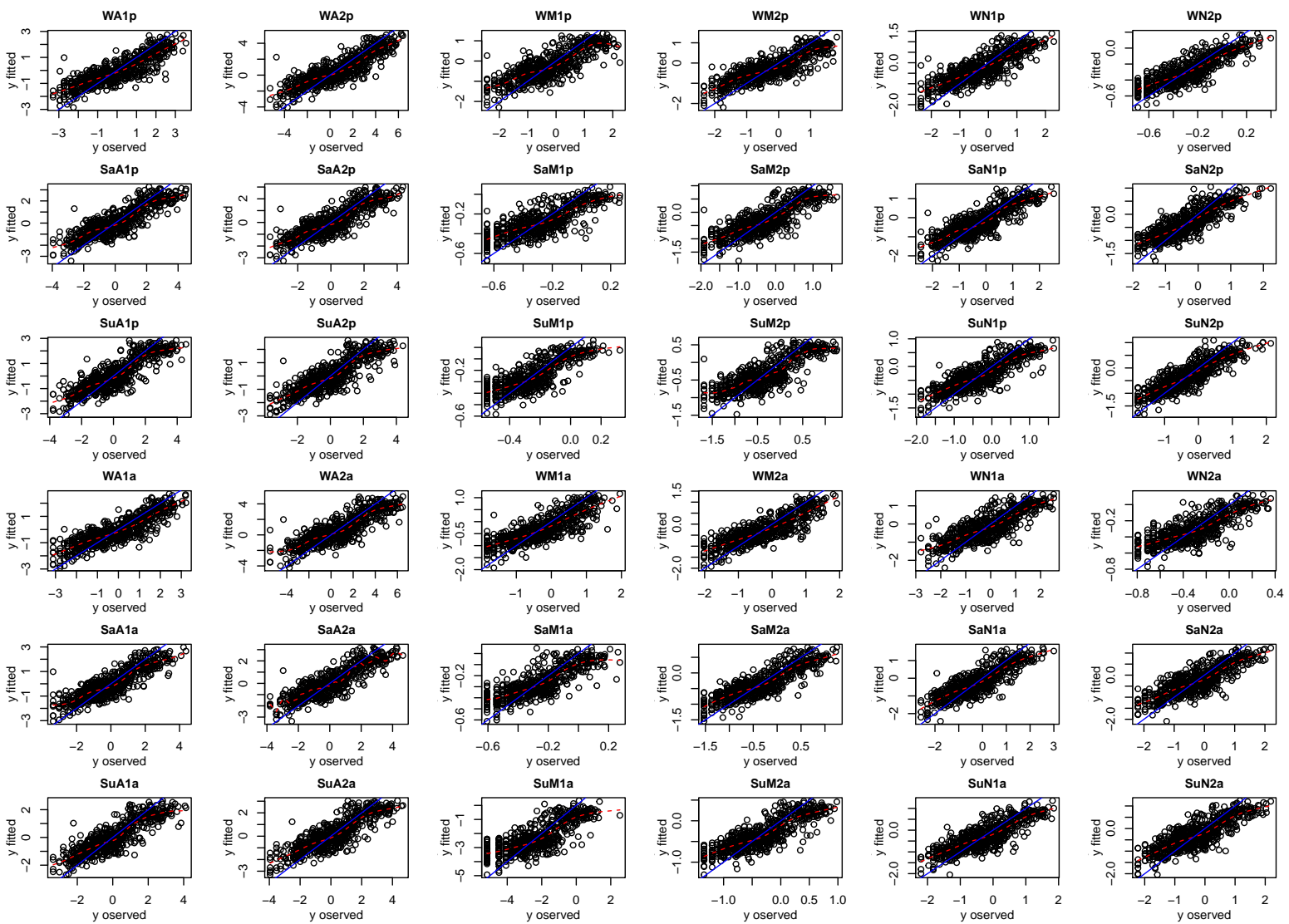


Figure E.6: GLM fitted vs observed values (Boxcox transformation)

E.3 GBM

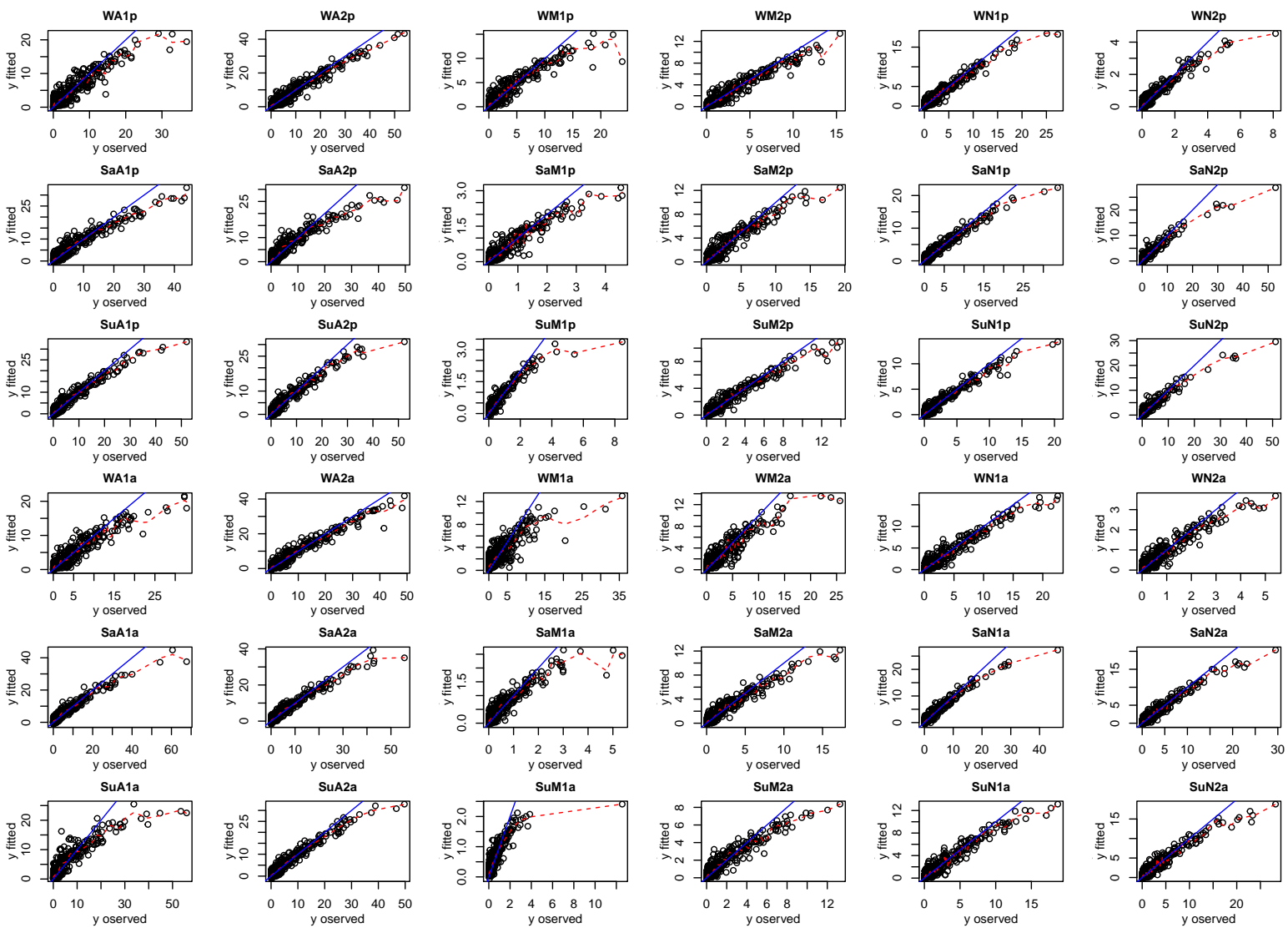


Figure E.7: GBM fitted vs observed values (No transformations)

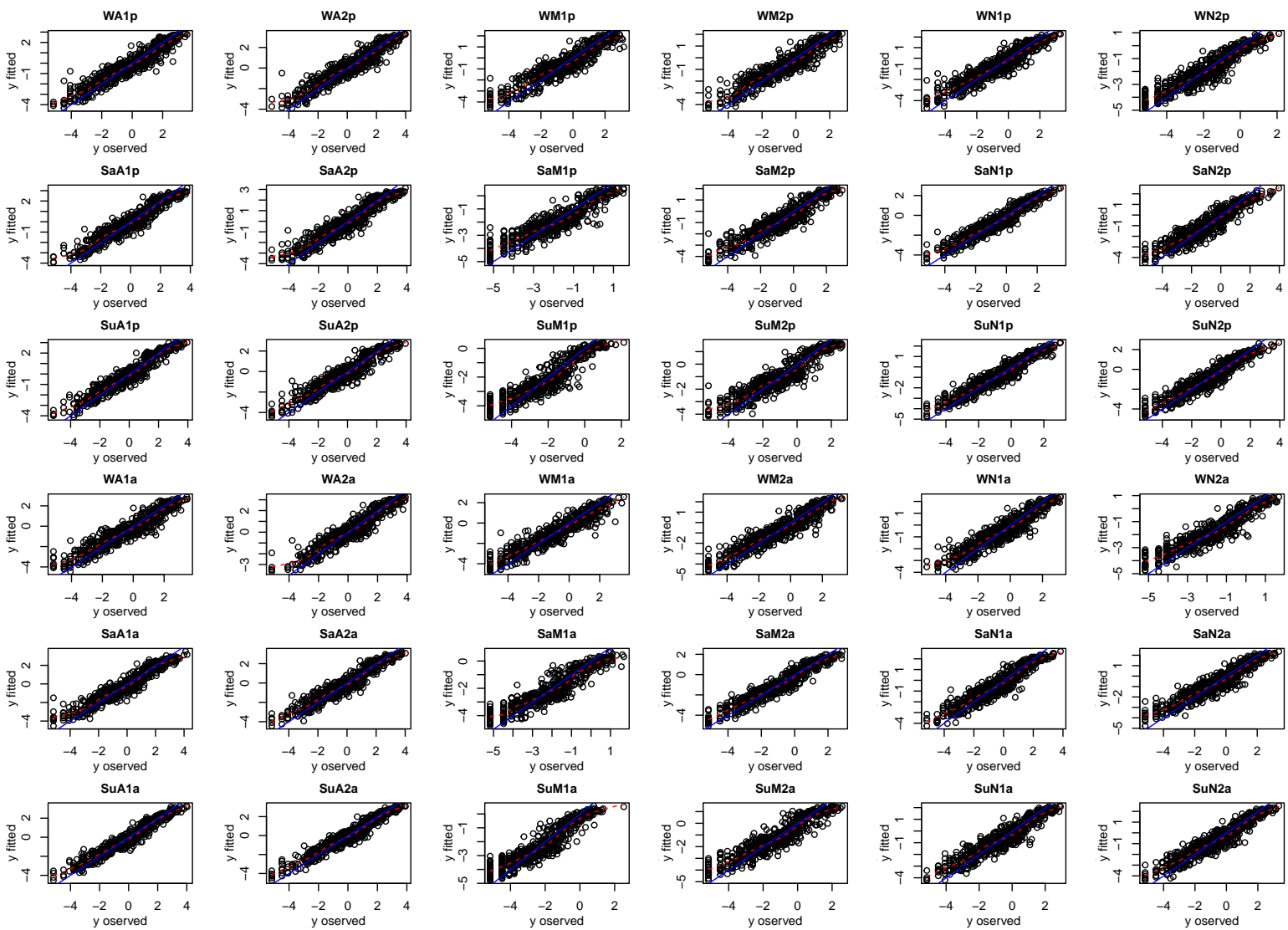


Figure E.8: GBM fitted vs observed values (Logarithmic transformation)

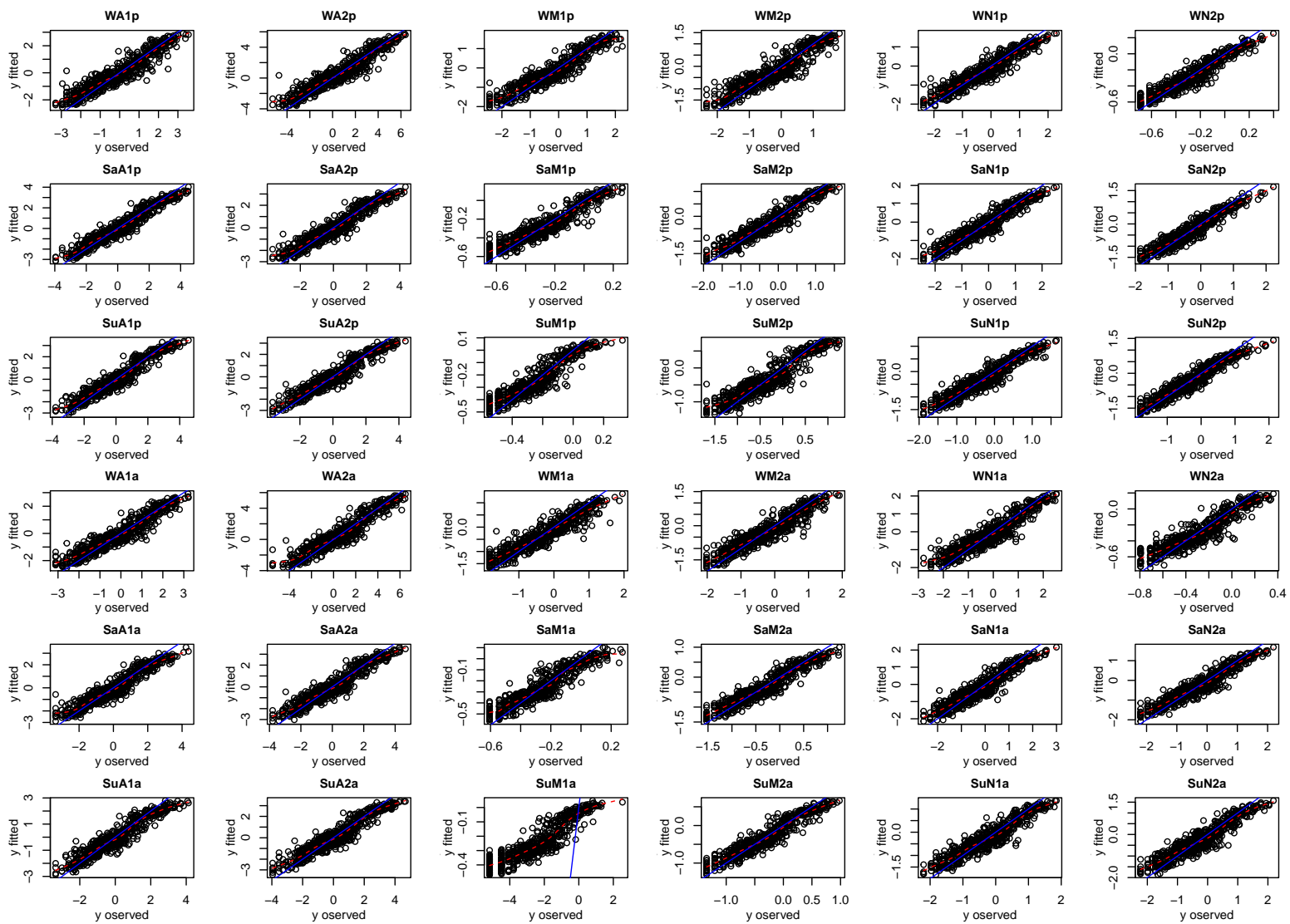


Figure E.9: GBM fitted vs observed values (Boxcox transformation)

Appendix F

Residual analysis

)

F.1 Stepwise regression

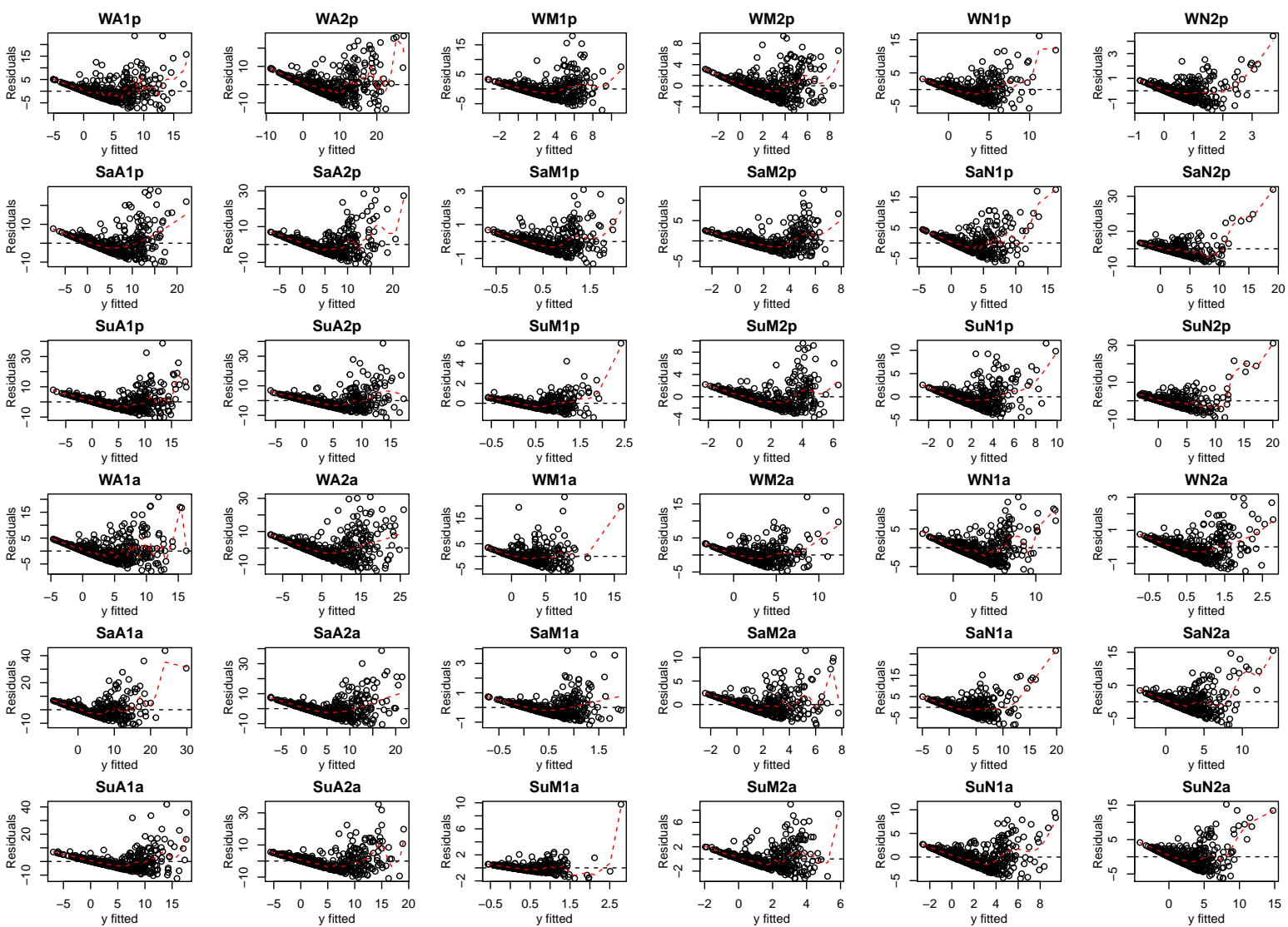


Figure F.1: Stepwise regression residuals vs fitted values (No transformations)

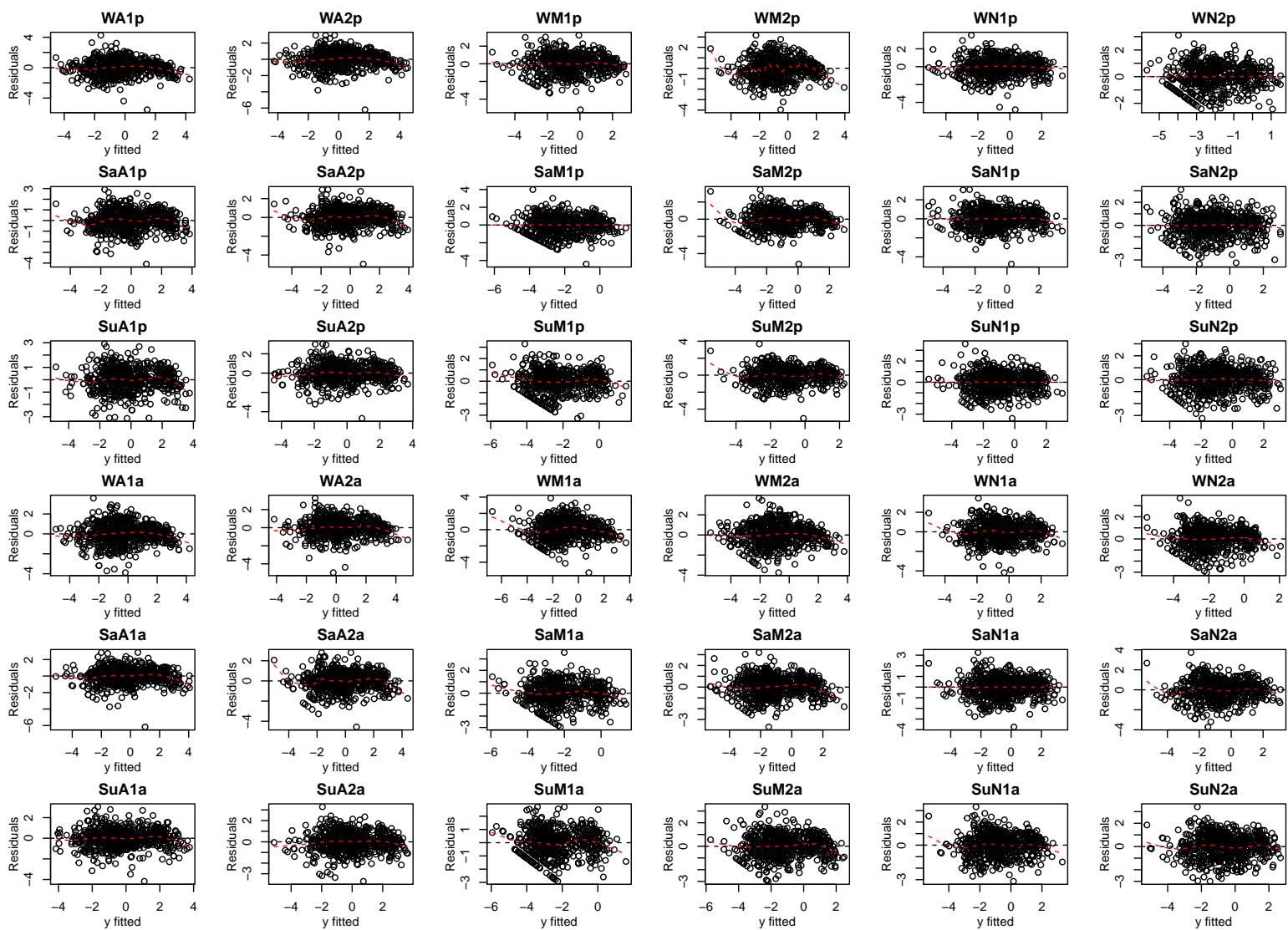


Figure F.2: Stepwise regression residuals vs fitted values (Logarithmic transformation)

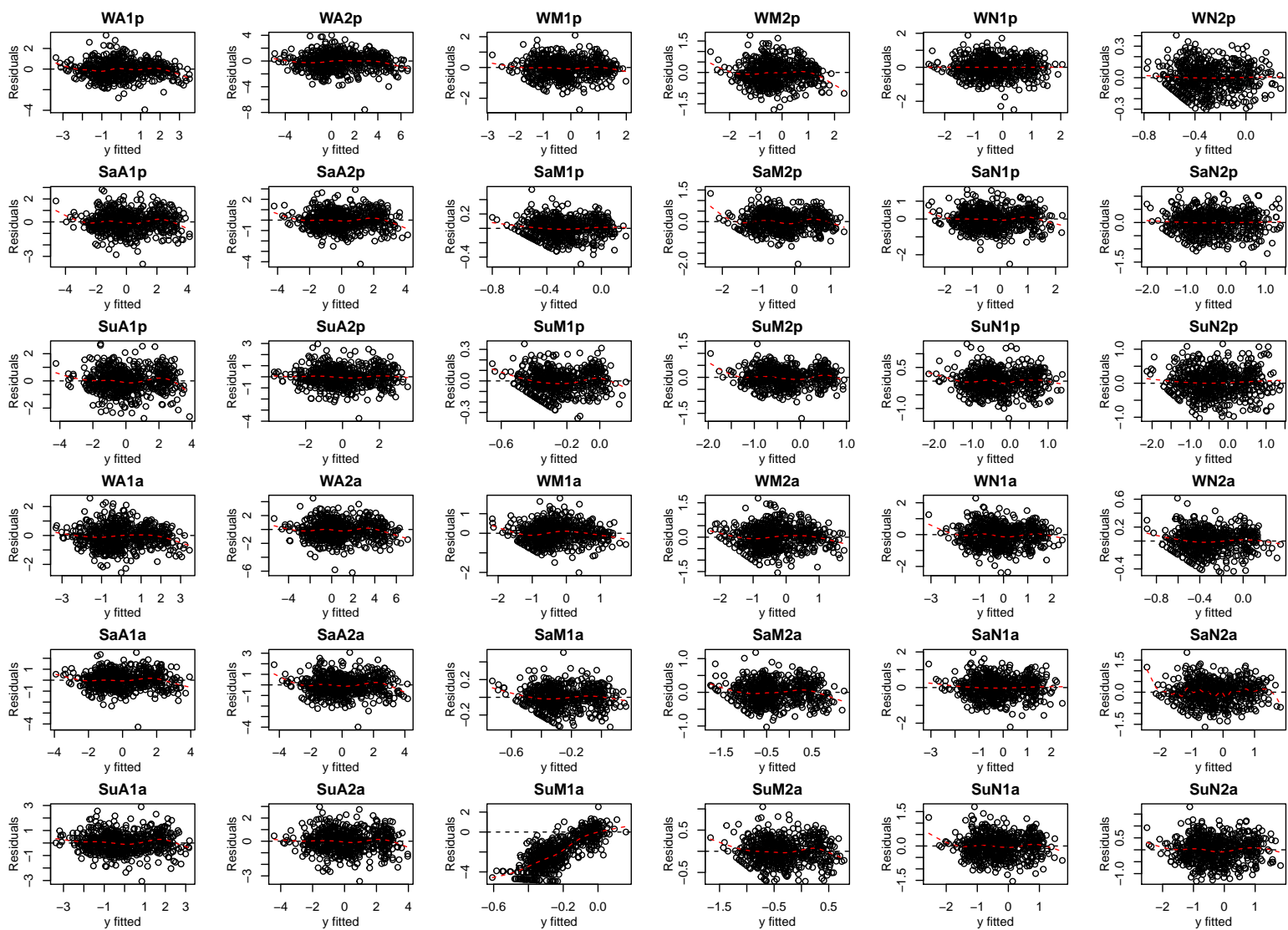


Figure F.3: Stepwise regression residuals vs fitted values (Boxcox transformations)

F.2 GLM

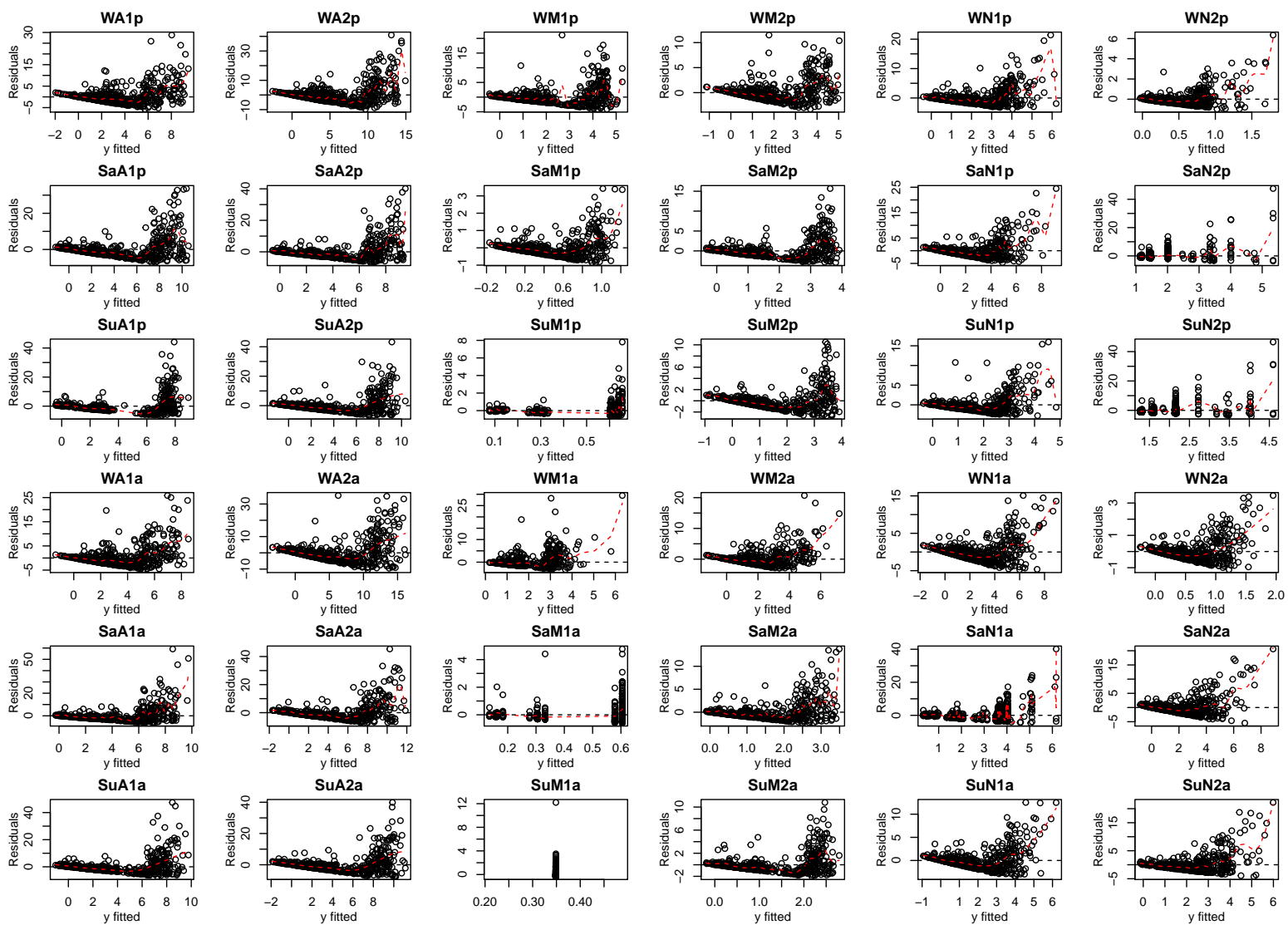


Figure F.4: GLM residuals vs fitted values (No transformations)

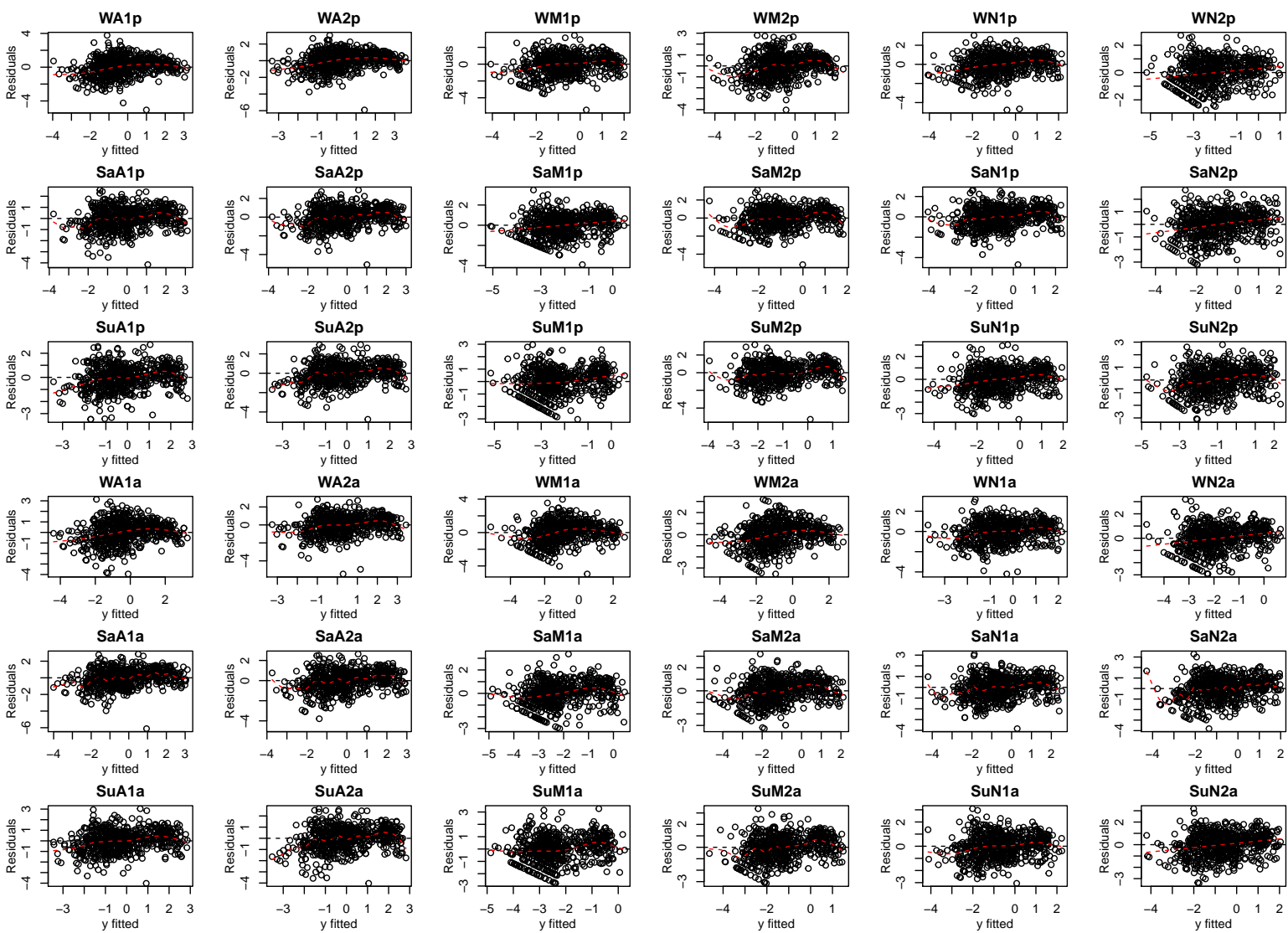


Figure F.5: GLM residuals vs fitted values (Logarithmic transformation)

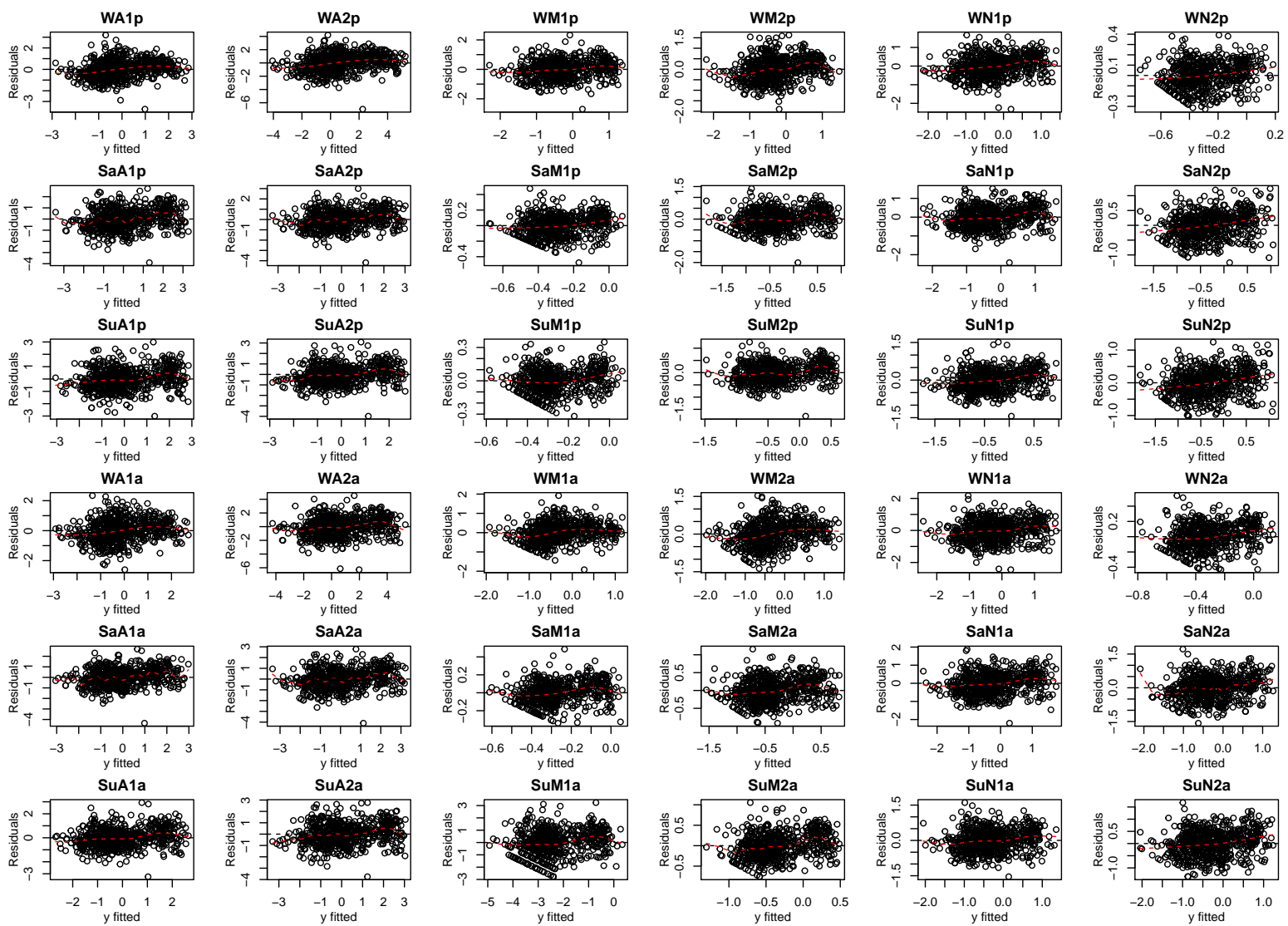


Figure F.6: GLM residuals vs fitted values (Boxcox transformations)

F.3 GBM

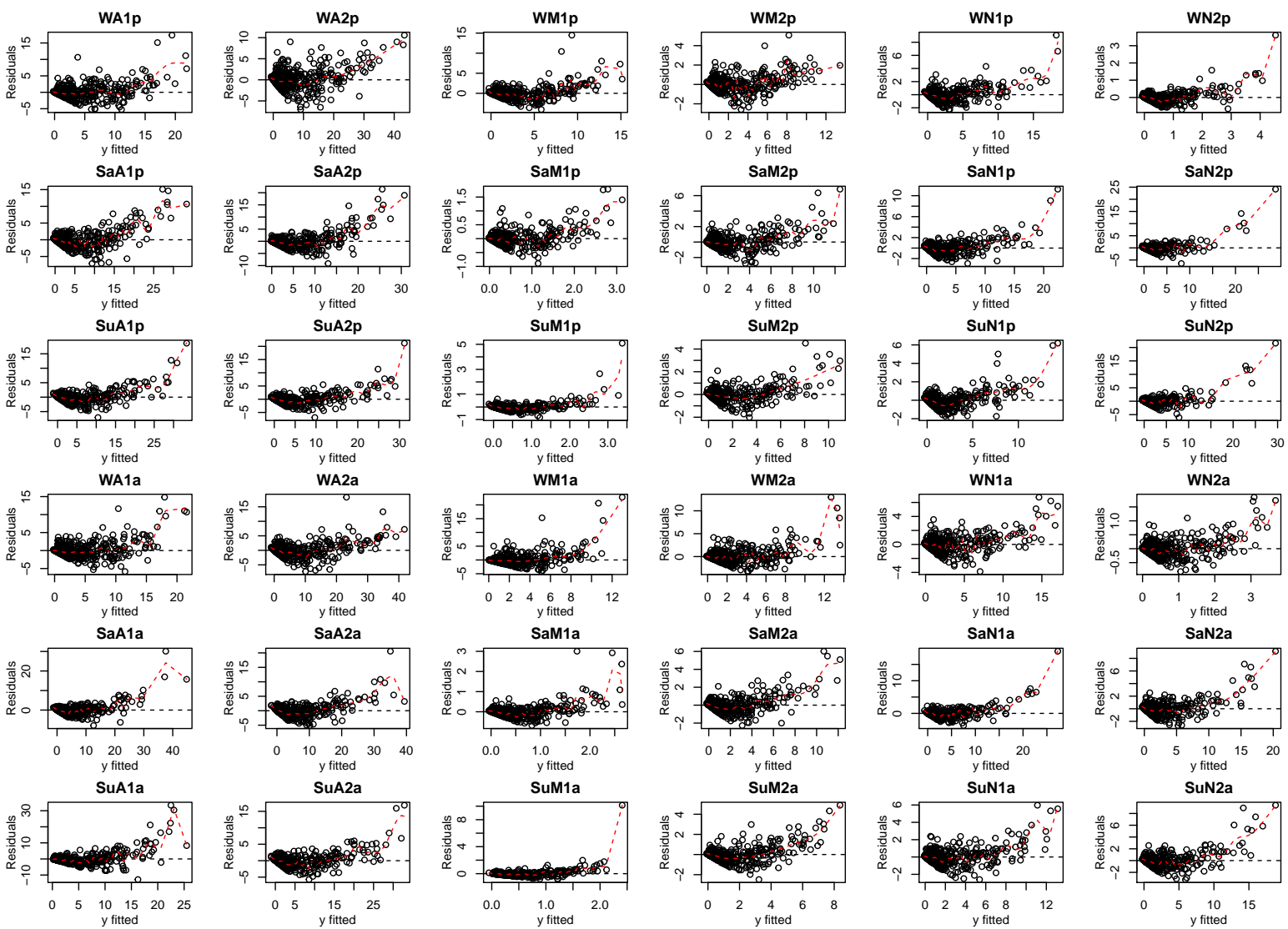


Figure F.7: gbm residuals vs fitted values (No transformations)

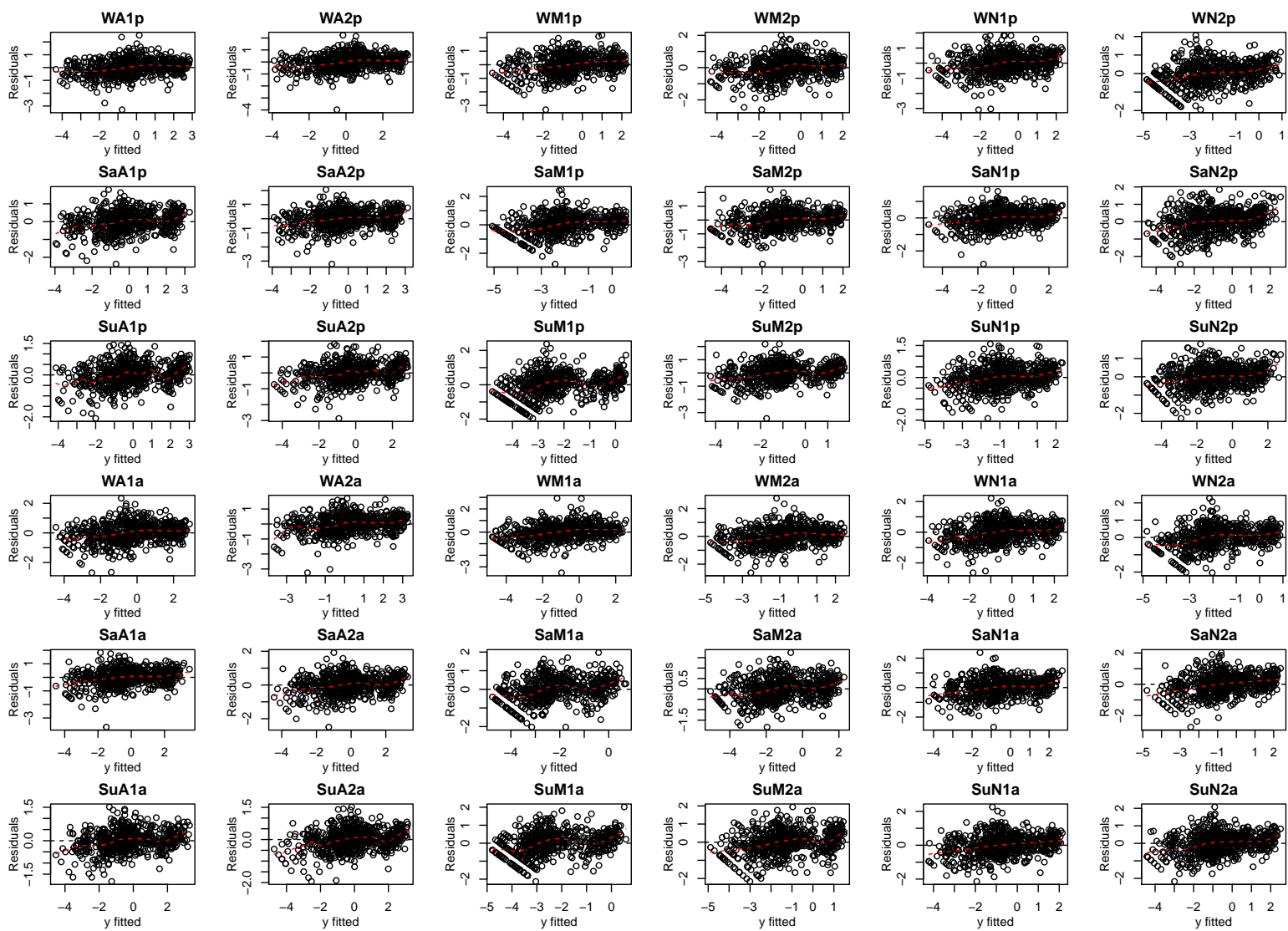


Figure F.8: gbm residuals vs fitted values (Logarithmic transformation)

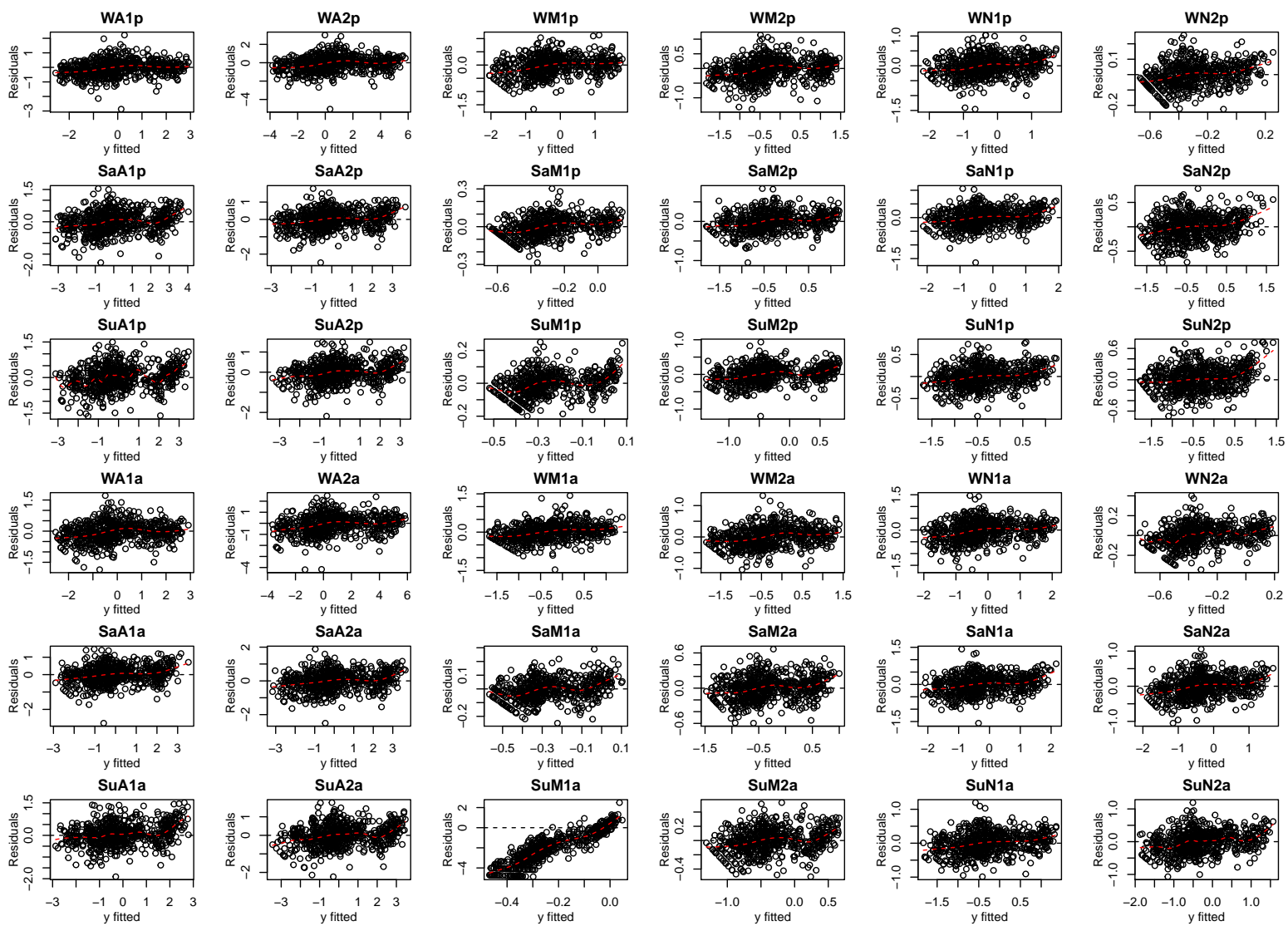


Figure F.9: gbm residuals vs fitted values (Boxcox transformations)

Appendix G

Predicted vs observed model values

G.1 Stepwise regression

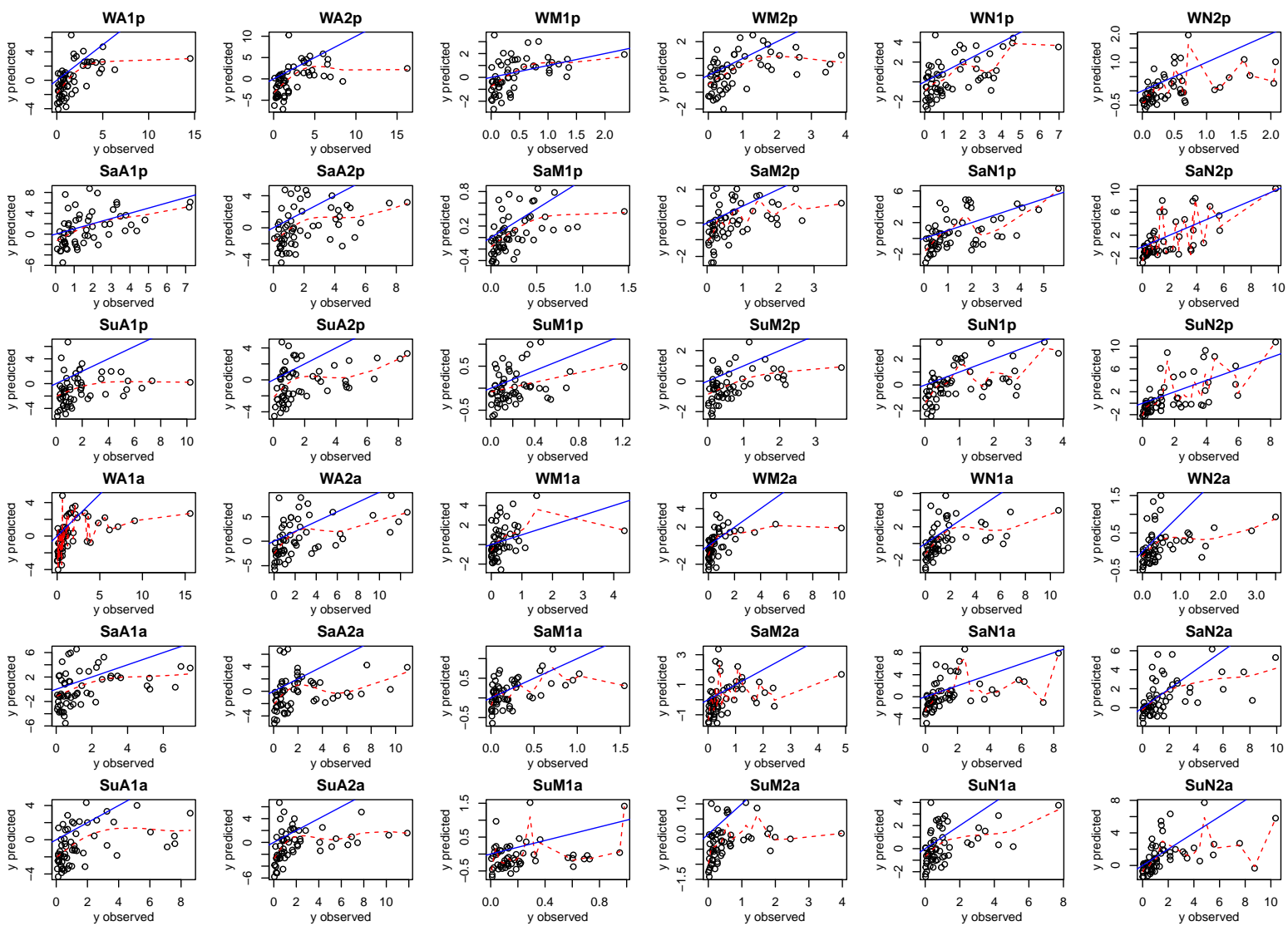


Figure G.1: Stepwise regression predicted vs observed values (No transformations)

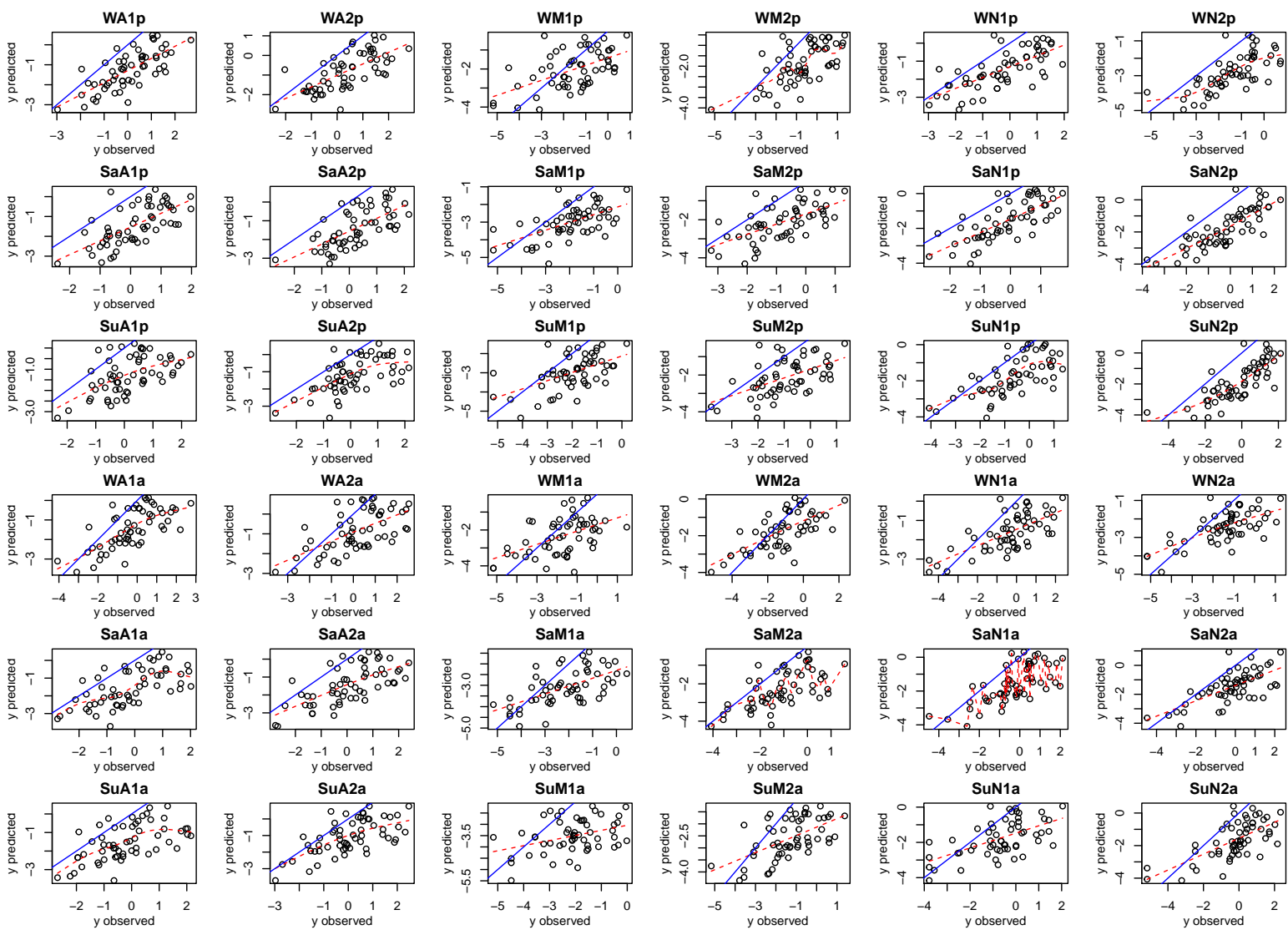


Figure G.2: Stepwise regression predicted vs observed values (Logarithmic transformation)

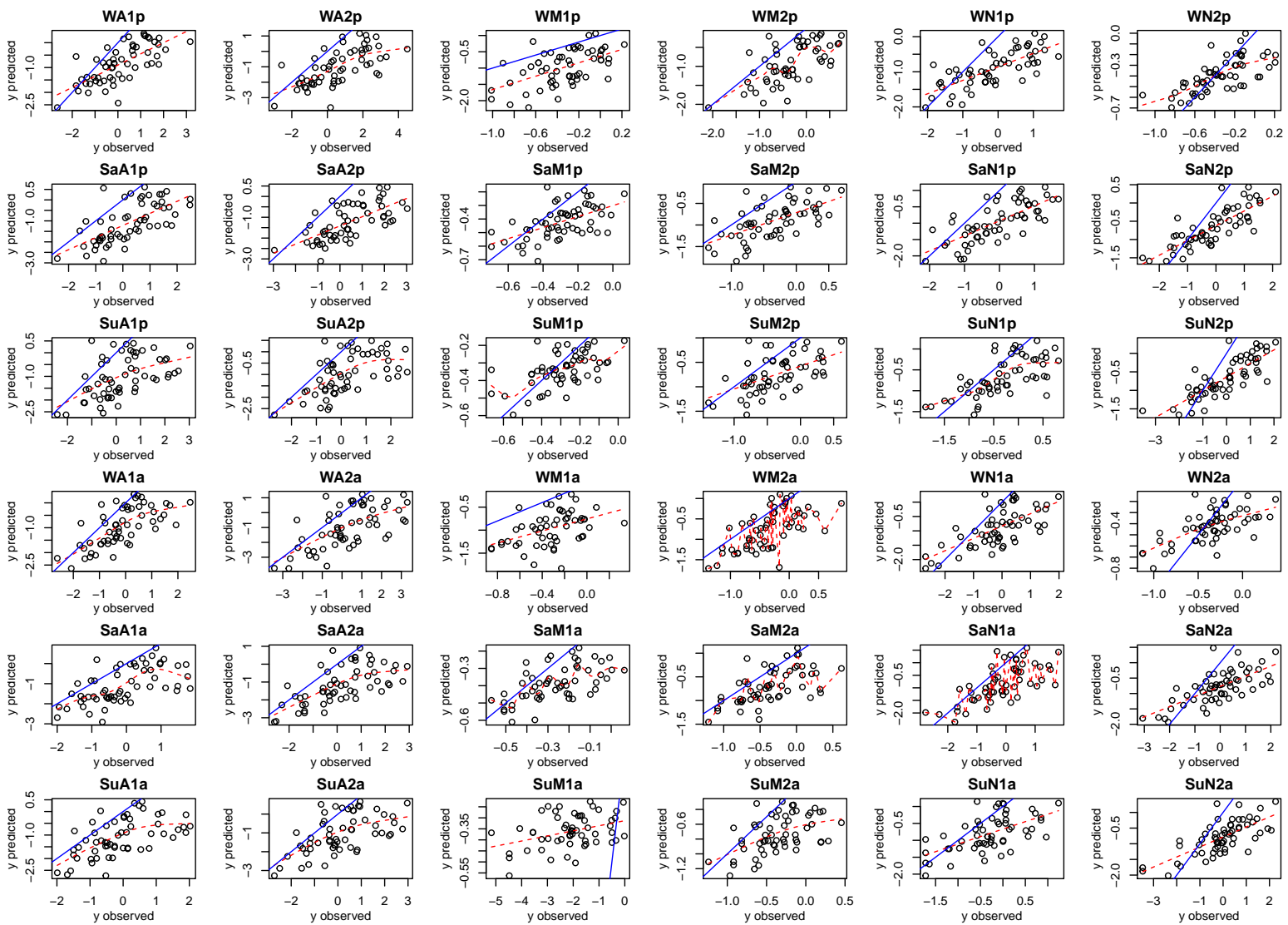


Figure G.3: Stepwise regression predicted vs observed values (Boxcox transformation)

G.2 GLM

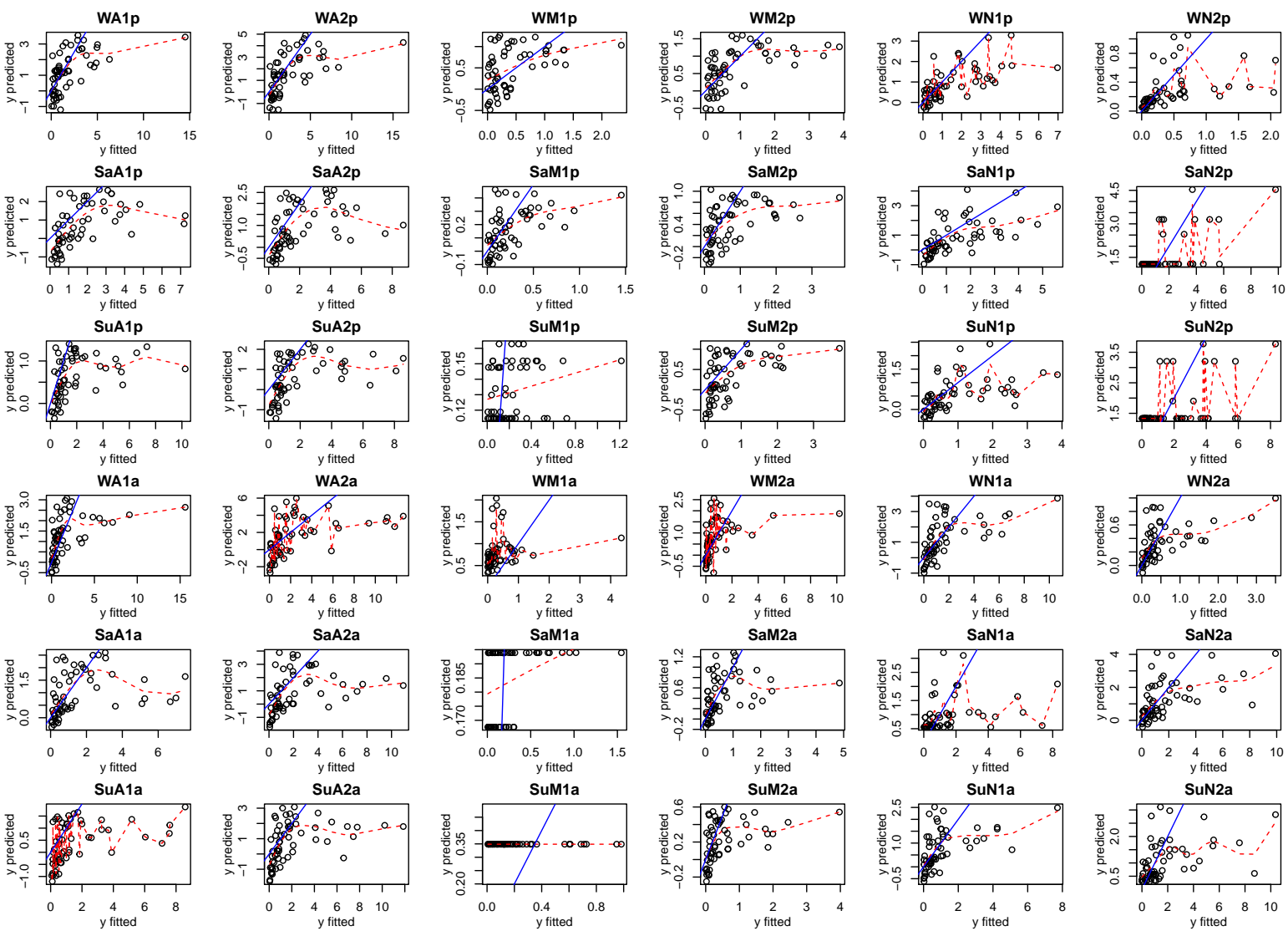


Figure G.4: GLM predicted vs observed values (No transformations)

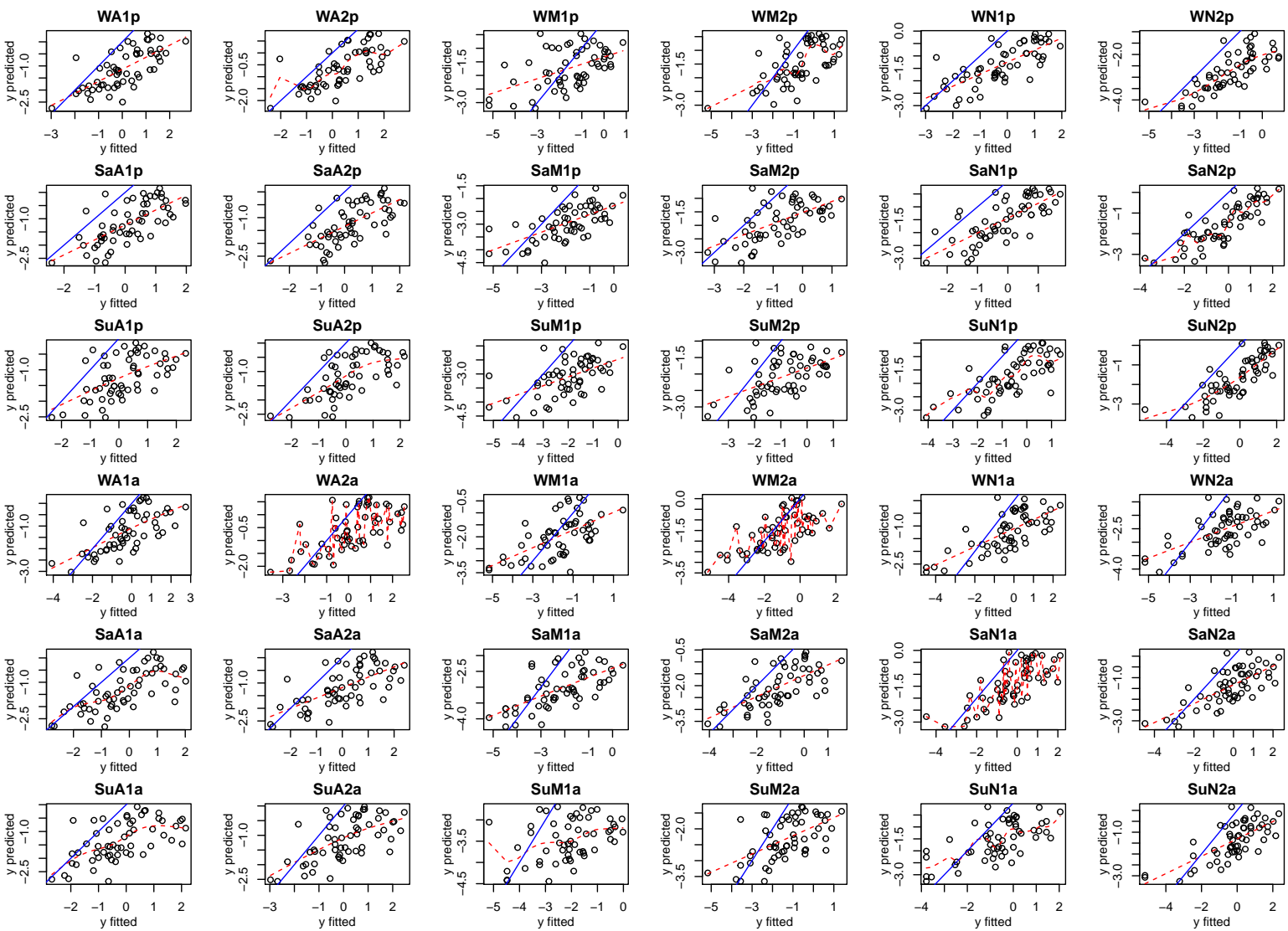


Figure G.5: GLM predicted vs observed values (Logarithmic transformation)

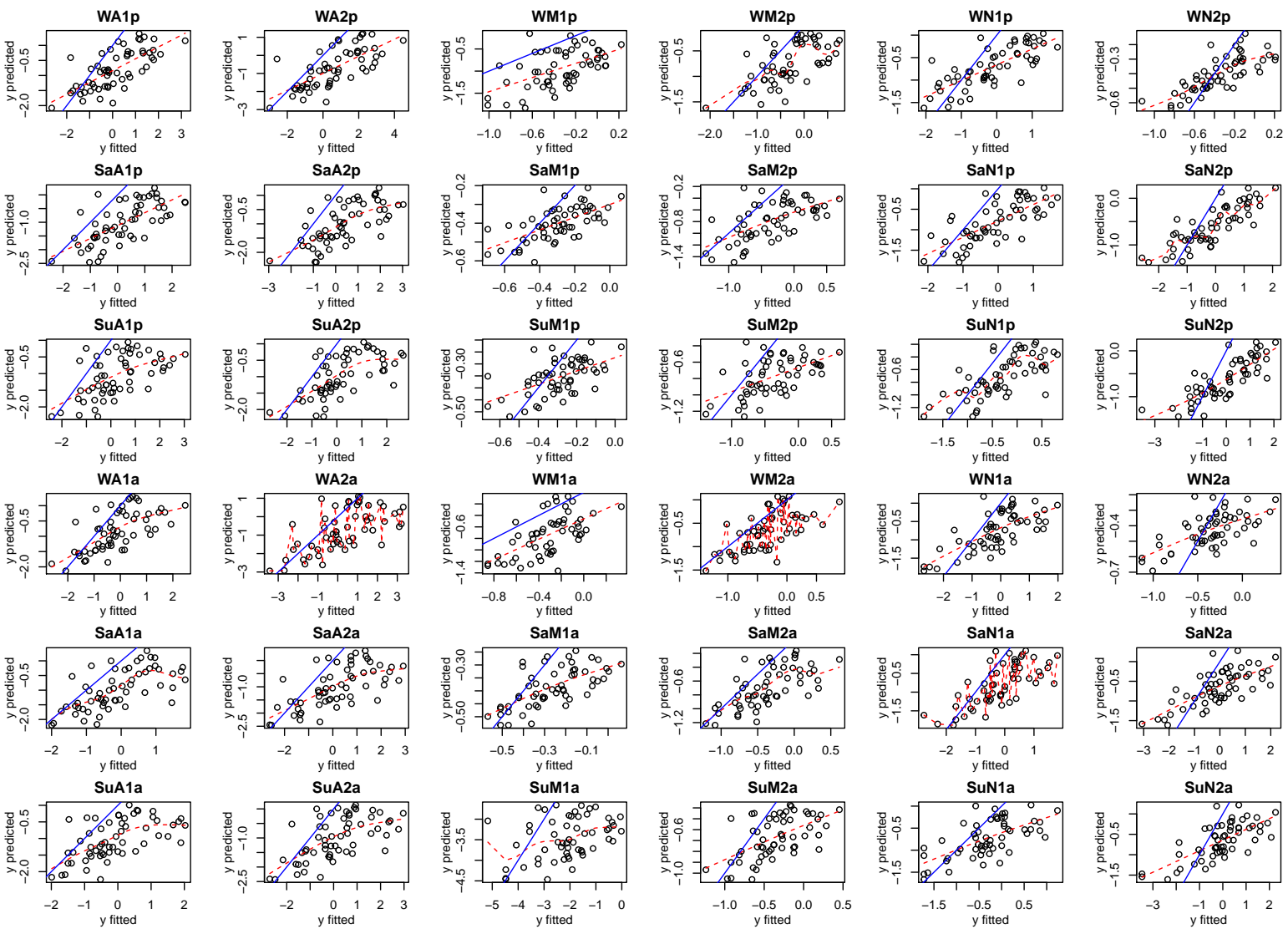


Figure G.6: GLM predicted vs observed values (Boxcox transformation)

G.3 GBM

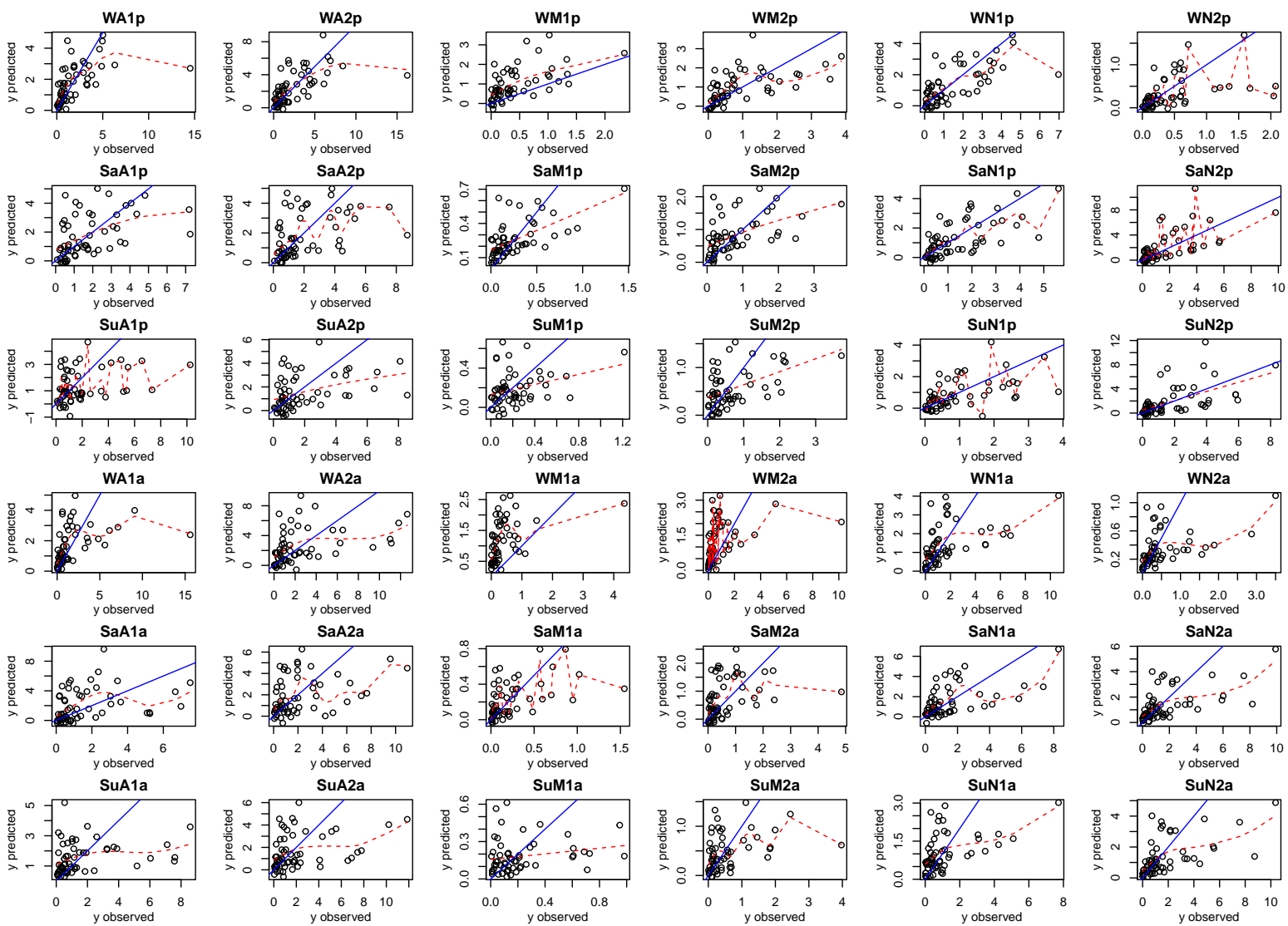


Figure G.7: GBM predicted vs observed values (No transformations)

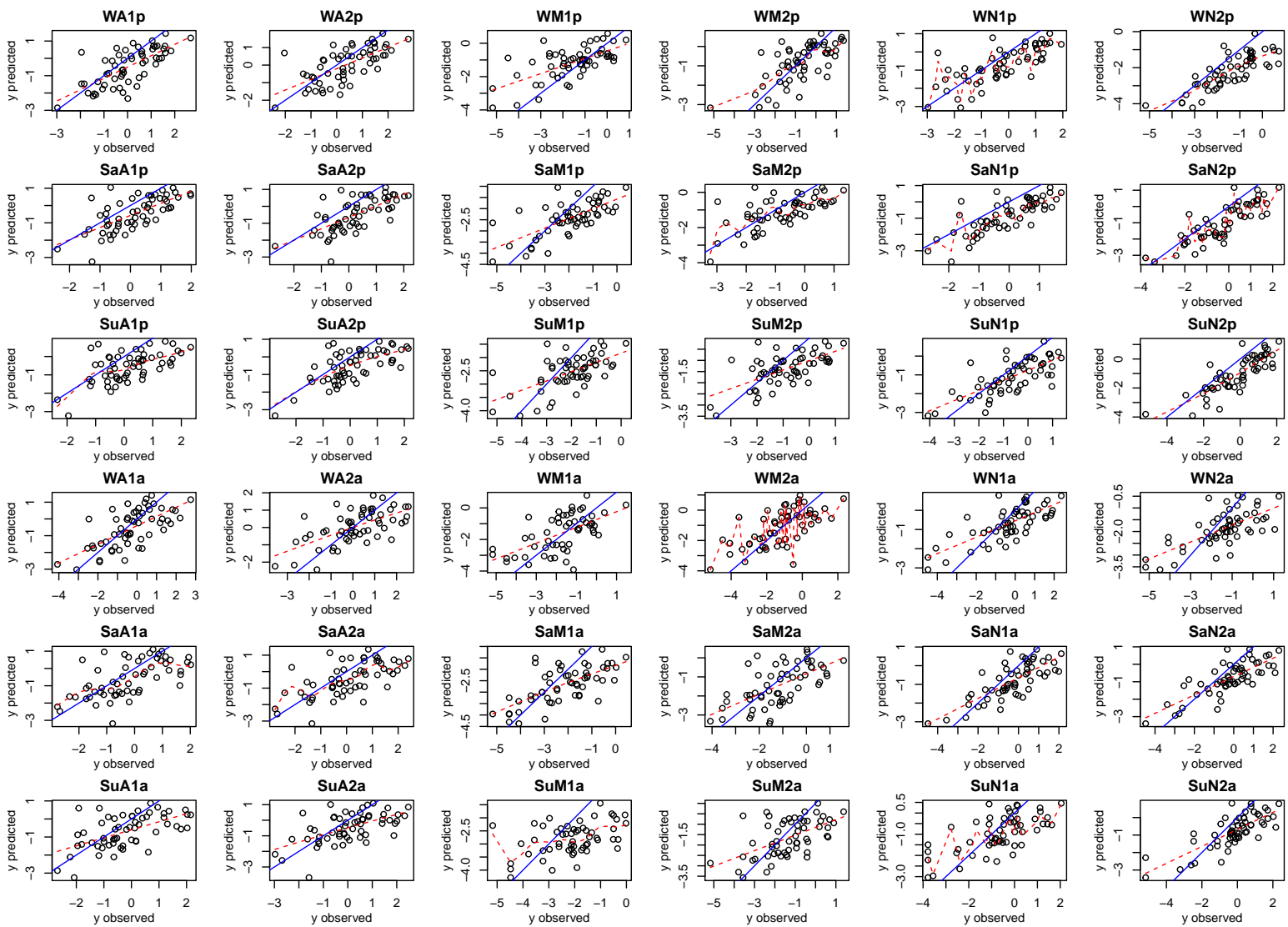


Figure G.8: GBM predicted vs observed values (Logarithmic transformation)

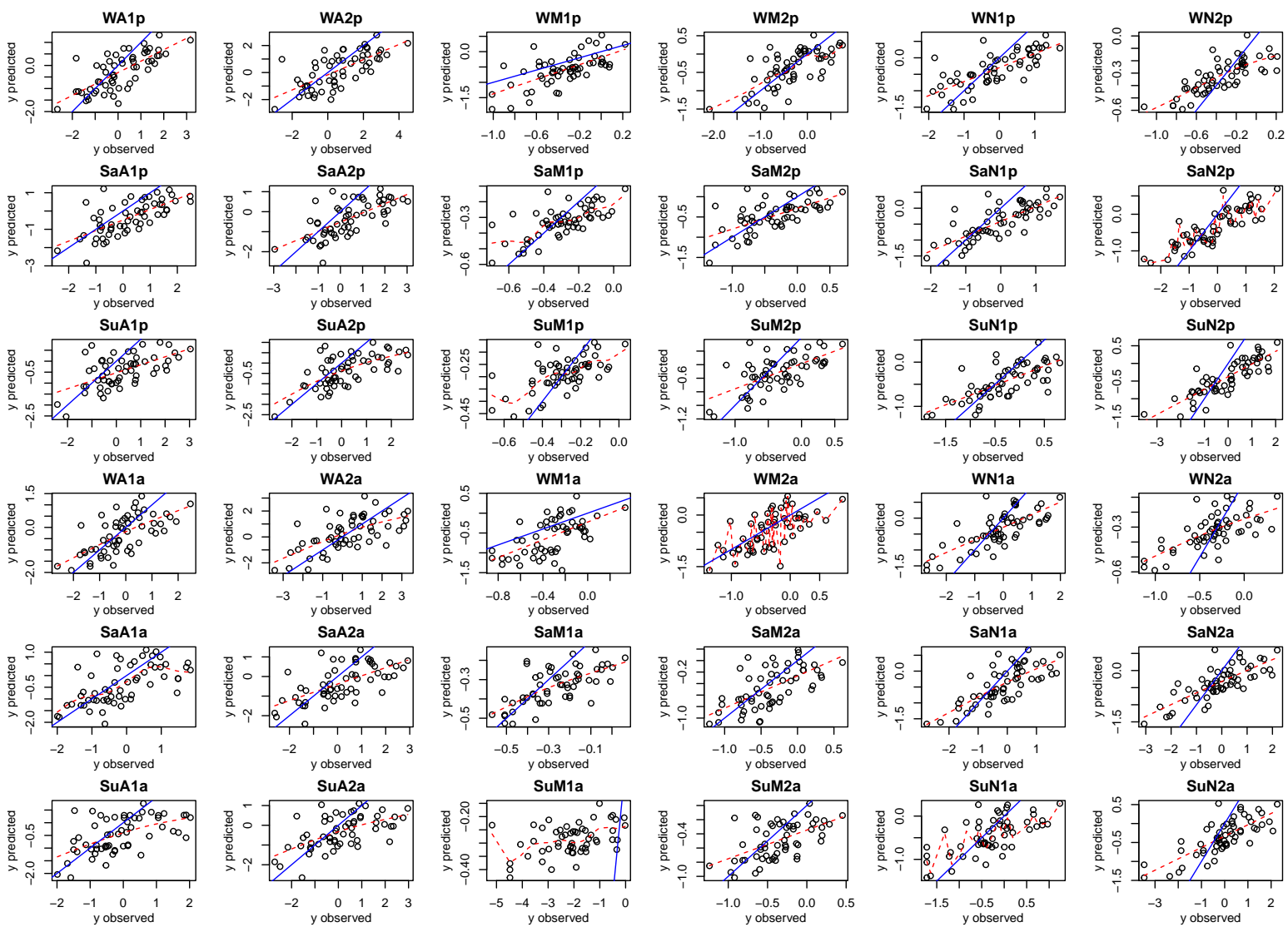


Figure G.9: GBM predicted vs observed values (Boxcox transformation)