BUILT ENVIRONMENT FACTORS AFFECTING BIKE SHARING RIDERSHIP: A DATA-DRIVEN APPROACH FOR MULTIPLE CITIES

David Duran-Rodas
Technical University of Munich
Arcisstrasse 21, 80333 Munich
E.mail: david.duran@tum.de

Emmanouil Chaniotakis
Technical University of Munich
Arcisstrasse 21, 80333 Munich
E.mail: m.chaniotakis@tum.de

Constantinos Antoniou
Technical University of Munich
Arcisstrasse 21, 80333 Munich
E.mail: c.antoniou@tum.de

Word Count: 7054 words + 1 table$\times$ 250 = 7304 words

Submission Date: August 1, 2018

ABSTRACT

Bike sharing has been found to present environmental, economic and social benefits. Identification of factors influencing ridership is necessary for policy–making, as well as when examining transferability and aspects of performance and reliability. In this work, a data–driven method is formulated to correlate arrivals and departures of station–based bike sharing systems with built environment factors in multiple cities. Ridership data from stations of multiple cities are pooled in one data-set regardless of their geographic boundaries. The method bundles the collection, analysis, and processing of data, as well as, the models' estimation using statistical and machine learning techniques. The method was applied on a national level in six cities in Germany, and also, on an international level in three cities in Europe and North America. The results suggest that the models' performance did not depend on clustering cities by size but by the relative daily distribution of the rentals. Selected statistically significant factors were identified to vary temporally (e.g. nightclubs were significant during the night). The most influencing variables were related to the city population, distance to city center, leisure-related establishments and transport related infrastructure. This data–driven method can help as a support decision-making tool to implement or expand bike sharing systems.

Keywords: bike sharing, built environment, open-source, multiple cities comparison

INTRODUCTION

Bike sharing is defined as the shared use of a bicycle, where a user accesses a fleet of bicycles offered on public space (*1*). It is part of the shared economy social–economic phenomenon, where individuals or organizations prioritize use over ownership of items (*2*). Bike sharing systems have a long history, with the very first system launched in 1965. Its deployment was in Amsterdam with fifty free and unlocked bicycles. Theft and vandalism led to a coin–deposit system, also not successful, mainly due to the user's anonymity. Nowadays, information and communications technology (ICT) enables wireless pick–up, drop–off, and a real–time GPS tracking of bicycles (*3*), which lead to the widespread of bike sharing to more than 1,600 cities around the world (4). (4). Europe and Asia are the continents with the majority of bike sharing systems worldwide. In 2015, China presented the biggest fleet in the world with 753,508 bicycles, followed by France with 42,930, and Spain with 25,084 (*4*). Categorization of bike sharing systems can be defined by the use of stations or not: (a) station-based (SBBS), b) free-floating (FFBS) and c) a mix of the two (*5*).

The wide deployment and observed growing trends of bike sharing can be attributed, among others, to its associated social, economic and environmental benefits. These are related to creating a larger cycling population, cost savings, increasing transit use, reducing greenhouse gases, decreasing congestion, creating environmental awareness, improving public health, among others $\big($a comprehensive review of benefits attributed to bike sharing can be found in (*3*), (*6*) and (*7*)$\big)$. However, not all systems were deployed successfully. Some were perceived as a public nuisance or were misused and vandalized (*8*). Possible reasons for a system failure were bicycles' poor quality, lack of funding, oversaturated market, delayed expansion, inconvenient system design, unfair fares, low political support (*8, 9*).

The identified benefits strongly suggest the necessity to further increase the use of bike sharing systems and to enable their deployment in more cities. At the same time, the unsuccessful deployment of some projects makes the examination of the factors that affect ridership and system reliability rather imperative. These two needs have been the driving force for a high number of studies on the influencing factors that affect the bike sharing usage (e.g., built environment, socio-demographic characteristics, system settings). Most studies analyze the influencing factors in a) multiple cities, with each city considered as one observation (*10, 11*) or b) single city at a local (station) level, where one city is analyzed and observations are based on an area of influence, e.g. near stations (*12–15*).

The multiple–cities approach suffers an exclusion of varying characteristics within a city, which provides an indication of how the system should be structured to enable a successful deployment. Conversely, the station level approach is performed in a single city and is bounded by the urban settings examined. The main issue with this approach is that it does not examine the system's transferability but the ridership within a city.

Aiming at overcoming the above–discussed drawbacks of existing approaches, this paper contributes to the related literature, by focusing on the investigation of bike sharing systems as one entity regardless of the city they belong to. We present a multiple cities data-driven approach focusing on the comparison of general built environment characteristics on a station–level influencing bike sharing systems beyond geographic boundaries. As such, the data used for each city is pooled in one complete data-set, where each observation refers to one station's area of influence (as defined in the Method Section). The influencing factors chosen to be investigated describe the characteristics of the built environment (guided by the high influence found in the majority of

previous studies).

The second main contribution of this study lies upon the modeling techniques used for the most influencing factors selection. As discussed in the Related Work Section, in most cases a predetermined set of factors is used. This set is hypothesized to contribute to a successful deployment of bike sharing, and thus, could omit possible patterns revealed by an alternative selection approach. In this study, a data–driven approach is followed to allow the discovery of factors that might not be commonly addressed. This is done by using different linear and non–linear modeling techniques which are evaluated upon modeling performance criteria such as goodness–of–fit, information criteria, and (cross–)validation.

Two applications of the methods discussed are included: a) a national application in six cities in Germany, Europe and b) an international application in three cities from Europe and North America with similar urban characteristics. The first application intends to illustrate the performance of different modeling techniques, while the second intends to illustrate the applicability of the methods in an international setting, while taking into account seasonality. In both cases, different validation techniques are exercised with very positive results. All the factors are defined using open–source data and by the derivation and deployment of an automated feature creation methodology.

RELATED WORK

The spatial-temporal factors influencing historical rentals of SBBS systems have been studied in cities all over the world with different sizes. The resulting factors have been compared between cities with factors within a city scale (*10, 11*) or in a single city on a station scale (*12–19*). The modeling approach followed in most of the cases above can be summarized as a) the model estimation method used is mainly a linear regression using ordinary least squares (*13, 15, 17–19*); b) the dependent variable is the logarithm of the number or rates of arrivals and departures (*12, 13, 15, 17, 19*); c) the model assessment is usually performed using the indexes: log-likelihood (LL), R^2 and AIC-BIC.

Regarding the multiple cities approach, De Chardon et al. (*10*) studied the trips per day per bicycle (TDB) in 75 SBBS systems in Europe, Israel, United States, Canada, Brazil and Australia. They used a robust regression to build the model with the logarithm of the TBD as the dependent variable. The resulting influencing variables were the operator's attributes, the compactness, the weather, the transportation infrastructure, as well as system–related characteristics, such as helmet requirement and the number of docks at stations. Faghih-Imani et al. (*15*) aggregated to an hourly value arrivals and departures in Barcelona and Seville into a Sub-City District level. They correlated both cities separately the of the logarithm of the dependent variable linearly to sociodemographic and socioeconomic variables and Points Of Interest (POI). Barcelona and Seville presented a similar pattern where the common influencing POIs were related to business, leisure, and restaurants.

On the other hand, considering a single city approach, Dung Tran et al. (*14*) developed models for bike sharing in Lyon using rain stations, restaurant, cinema and embankment roads, altitude, among others. They also found that the population density showed a positive effect in the morning and the number of jobs with a positive impact in the afternoon. Faghih-Imani and Eluru (*12*) correlated the hourly arrivals and departures for one month in the SBBS "CitiBike" in New York with temporal, spatial and weather variables. They concluded that the fit of the model improved significantly by adding temporally and spatially lagged dependent variables. The length

of bicycle routes, the presence of subway stations, the area of parks on weekends, and the number of restaurants increased the usage of the system, while the length of railways decreased it.

Activities that are associated with bike sharing are commuting and leisure (*20*). These activities are vastly related to the built environment. Thus, many studies have examined its relationship to bike sharing upon the transport infrastructure, Points Of Interest (POI) and the land use categories (Table 1).

TABLE 1: Built environment factors most influencing bike sharing historical data

CATEGORY	VARIABLE	REFERENCES								
		(*10*)	(*12*)	(*16*)	(*17*)	(*13*)	(*18*)	(*14*)	(*15*)	(*19*)
Transport	Cycling infrastructure	✓	✓							
	Railways length		✓							
	Subway stations		✓	✓		✓		✓	✓	
	Rail stations							✓		
POIs	Universities					✓	✓			✓
	Student residence							✓		
	Restaurants		✓		✓		✓	✓		
	Cinema							✓		
	CBD				✓		✓			
	Number of business				✓		✓		✓	
Land use	Parks		✓							
	Residential land use			✓						
	Parking land use			✓						
	Bodies of water				✓					

To the best of the authors' knowledge, there is no data-driven method to measure built environment variables by assigning them automatically different types of indicators as quantity in the area of influence of the stations or proximity to the stations. Contrary to De Chardon et al. (10), we analyzed multiple cities, but on a station scale. Also, instead of comparing the influencing factors of multiple cities (*15*), we model the cities together. Finally, in the literature review, there was not a comparison of different linear and non–linear modeling techniques to define which technique fits better the bicycle usage and the influencing built environment.

METHOD

The proposed method aims at building models automatically in different temporal scales to identify the built environment variables that influence the historical rentals of SBBS systems in multiple cities. The method goes through three main components: 1) automated data collection, 2) automated data analysis and processing, and 3) automated model building and selection of the modeling technique with the better fitting results, and automated selection of the most influencing variables (Figure 1).

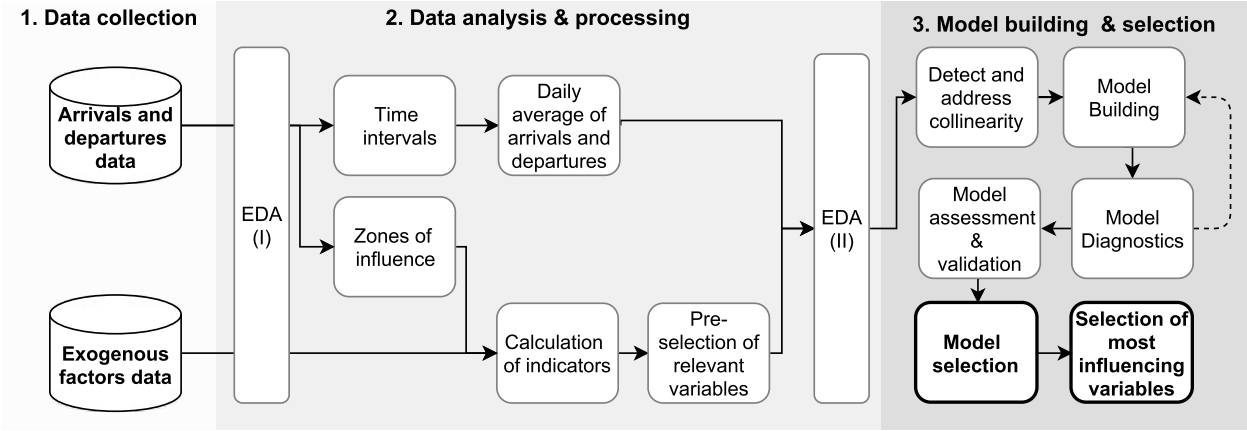


FIGURE 1: Methodological framework

Data collection, analysis, and processing

Data collection is performed on historical arrivals and departures from bike sharing systems (dependent variables) and the built environment (independent variables) in multiple cities. The independent variables are points, lines, and polygons of the built environment: e.g., points of interest, public transport stations, railways, roadways, waterways, land use, and natural features.

Ridership Data

The historical ridership data (in terms of arrivals and departures to and from a station) are explored to define time intervals to build models independent of time. This is performed to allow homogeneity in terms of dependent variables and to correct the effects of the time of the day. A clustering analysis is carried out to determine which days of the week illustrate significantly different ridership patterns. In each cluster, different periods are identified based on the hourly distribution of the rentals based on peak and off–peak times. The cumulative ridership variable (dependent) and built environment variables are aggregated on a spatial scale, based on zones of influence. These zones are defined as the maximum area of influence that an individual is willing to walk to reach a bike-sharing station. Their boundaries are defined as the intersection of the Thiessen polygons of the stations, human-made and natural barriers and a buffer circumference from the stations representing the maximum walking distance [200 to 400 meters (*10, 13, 14, 16–18, 21*)] that a station can attract or produce.

Built environment Data

Built environment data is downloaded from an open–source database on a city level. Each built environment variable is assigned two indicators in each zone of influence: 1) proximity-based indicators (minimum distance from a station to the examined spatial feature inside the zone of influence, and 2) quantity or presence of the variable in a zone of influence. The selection of the appropriate type of indicator is decided based on some basic hypotheses. Let v be a random (independent) variable used to describe a particular built environment distribution across observations. Also, let σ_v represent its standard deviation. A variable is defined as static if the standard deviation is smaller than a threshold t ($\sigma_v < t$). Under the above hypotheses, indicators will be introduced in the model as dummy variables indicating "presence" rather than quantity. A sensitivity analysis is carried out to determine the value of the threshold of the standard deviation. Only the variables

that are present in all the cities of the study are considered. These variables are explored to exclude those presenting inconsistencies or irrelevant to the influence of bike sharing ridership.

Finally, Pearson and Spearman correlation tests are carried out to determine the variables that are collinear. If two variables are collinear, the variable that influences more the rentals is considered (multiple regression models are estimated).

Model building and selection

Model building and selection is based on a sequential model definition and model validation process. The aim pursued is to build linear and non-linear models to identify the models that better fit the dataset, while a) being parsimonious without a substantial loss of their fitting performance, b) avoiding over-fitting and c) including variable selection or variable categorization for computational efficiency, given a large number of independent variables. Mathematical transformations of the variables are considered to handle heteroscedasticity or non–linearity issues and to improve the models. Most common transformations are the square root, logarithmic, inverse, exponential, arcsine (*22*), or Box-Cox transformation (*23*).

Model Structures

The model building techniques to be examined are stepwise Ordinary Least Square regression (stepwise OLS) (*24*), Generalized Linear Models (GLM) with a lasso selection technique (*25*), and Gradient Boosting Machine (GBM) (*26*).

Stepwise OLS for variables selection is chosen based on its wide use in the pertinent literature for similar types of problems (*11, 13, 15, 17–19*). The core of stepwise OLS is the multiple linear regression, which is iteratively used to build a model using an observations vector Y that is linearly related to a matrix X (independent variables) and ε residuals [$Y = X \cdot \beta + \varepsilon$ (*24*)]. Stepwise regression addresses the subset selection of a large number of k parameters. There are three types of stepwise selection procedures: 1) forward selection, 2) backward selection 3) both directions (*27*). The forward selection initiates with only the constant term (i.e., no parameters) and adds variables based on a comparison criterion. The backward elimination process, in contrast, starts with a full equation and excludes the uncorrelated parameters. The stepwise method in both directions sequentially adds or deletes parameters. It starts with a forward selection, but at each step, it can remove a parameter. Its advantage is if a non-significant parameter is included in the process, it might be eliminated later.

The selection of the parameters is based on criteria to compare the regression in each step. Commonly used criteria are the Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) (*24*). AIC (*28*) is defined as:

$$AIC = n * ln(MSE) + 2 \cdot k \tag{1}$$

where n denotes the number of observations, MSE the mean squared error and k the number of parameters. A direct implication of using AIC is that for two models with the same error, AIC would penalize the one with more parameters. However, the use of AIC tends to improve with a larger number of k parameters, thus it is commonly accused of being prone to overfit models selection. BIC (29) tends to control the overfitting of AIC. It is proportional to AIC but it uses a logarithmic factor for the effect that the number of variables has:

$$BIC = n * ln(MSE) + ln(n) \cdot k \tag{2}$$

Generalized linear models (GLM) are an extension of OLS. Usually, this family of models is based on the maximum likelihood estimation. GLM assume that the error ε 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 $[h(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i1} + ... + \beta_1 x_{ik}$ where, $h(\mu_i)$ is a function that links μ_i with the observation Y_i] (*24*). The least absolute shrinkage and selection operator (lasso) technique (*25*) shrinks the coefficients β increasing stability while retaining the best variables. Lasso assumes that $X_i j$ are standardized with a mean of zero and a standard deviation of 1. Then, it minimizes the sum of the squared differences between the observation and the linear regression $[\hat{\beta}_{lasso} = argmin_\beta(\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{k} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{k} |\beta_j|)]$. Models generated with lasso are easy to interpret. The selection of λ is calculated after a cross-validation test to select the λ that presents the smallest error (*27*).

Gradient Boosting Machine (GBM) is a machine learning algorithm that performs regression, classification, and ranking (*26*). It is a mix of boosting and gradient steepest descent. 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 $\hat{f}^1(x), \hat{f}^2(x), ..., \hat{f}^B(x)$ 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. 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 fit function. Finally, it combines all the trees to create a single model. A faster approximation to find the model is to consider a differentiable loss criterion that can be derived by numerical optimization. Regarding the loss function, a Gaussian function is used for numerical efficiency to minimize the squared error and the Laplace for minimizing the absolute error.

GBM is considered in the study because the dataset might fit better in a nonlinear model. It uses the input arguments: loss function, number of iterations, terminal nodes of each tree and shrinkage factor (30). A sensitivity analysis has to be carried out to determine these values. In addition to the resulting model, GBM provides a ranking list of the variables with their relative influence normalized to sum one hundred. To carry out a variable selection from the ranking list, mean square errors (MSEs) are calculated starting from highest ranked variable and then adding a subsequent variable until a non–significant difference of the MSE is present. GLM and GBM have shown high fitting performance in similar applications in the literature, e.g., (*31*).

Model building is carried out with a training set and the model validation with a testing set for each time unit and for the linear and non–linear regression models. After the models are built, two types of criteria are used to assess them: a) Indirect methods: lowest number of predictors, lowest Mean Square Error (MSE), lowest BIC, and greatest goodness of fit measures (R^2 and adjusted R^2 - R^2_{adj}); and b) Best validation results: selection of the model that adequately predicts the arrivals and departures on a validation dataset.

APPLICATION

The data–driven method was applied in two cases: 1) national level in six German cities and 2) international level with three cities in three countries in Europe and North America. The national level application provides evidence on the applicability and performance of the different model structures and estimation techniques, allowing for a more comprehensive evaluation of the impact that different techniques have to the identification of the factors affecting ridership. Germany has

been used as the national case, due to the bike sharing fleet (fifth largest fleet in the world; approx. 12,000 shared bicycles) (4) and data availability. The international level application builds on the first application and focuses on the extraction of conclusions for the application of the methods in an international comparison.

Multiple National Cities approach

This approach includes six German cities for the SBBS system "Call a bike" (*32*): Hamburg, Frankfurt am Main, Stuttgart, Kassel, Darmstadt, and Marburg. Arrivals and departures of the 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` on June 2017. The dataset included the rentals in fifty cities in Germany for approximately 3.5 years. The majority of the data, however, referred to the six selected cities, because of their high usage of bike sharing (>250,000 rentals in total, around 3.5 GB). In total, 10.5 million rentals were included in the dataset referring to the period between 01.01.2014 and 15.05.2017 (1232 days). Around 73% of the rentals referred to the city of Hamburg, followed by Frankfurt with 12%. Peaks were identified in the summertime (May to July). It is worth mentioning that Wednesdays and Thursdays showed the highest ridership, which was found to decrease during the weekends. Regarding the hourly distribution, there was a different trend between workdays and weekends (Figure 3). In workdays, there were two peak periods at 8:00 and at 17:00 (based on the median values). Figure 2 shows the spatial distribution of the intensity of the rentals. Each area represents the frequency of a station with the help of Voronoi Diagrams for better visualization.

Data analysis and processing

The rentals were clustered into days of the week using Pearson correlation analysis. The three resulting clusters were workdays, Saturdays, and Sundays. Arrivals and departures were aggregated into time intervals representing peak and off–peak periods in the morning, afternoon and night (Figure 3). The built environment variables were downloaded from the collaborative open–source dataset OpenStreetMaps (*33*). Unclassified roads and a selection of variables, which were found to be inaccurately positioned or irrelevant, were excluded (e.g., vending machines, wastebaskets). The distance to the city center was also considered as a built environment variable since it was present in the literature review. For the zones of influence, a 300 meters buffer ratio was used as being the most common value used in the literature. Four indicators were assigned to around 200 types of spatial features (around **800** spatial variables). A threshold value selected of $SD = 5$ was selected after a sensitivity analysis to determine if the indicator of a variable is related to the quantity or presence. A total of **194** variables were examined with **144** non–collinear variables to be selected after Pearson and Spearman correlation tests. A threshold value of 0.7 was considered as explained in Zhao et al (*11*).

Model building, diagnosis and validation

Aiming at examining applicability and performance of the different model structures and estimation techniques, all methods discussed in the Methods section were used (OLS, GLM with lasso and GBM). In all cases, the relationship between arrivals and departures to 144 non–collinear built environment variables was examined. The city population was used to weight ridership (for different cities' sizes) (*34*). Apart from model fitting and model diagnostics, model validation was performed by dividing ridership data into a training set including the zone of influence of 5 cities

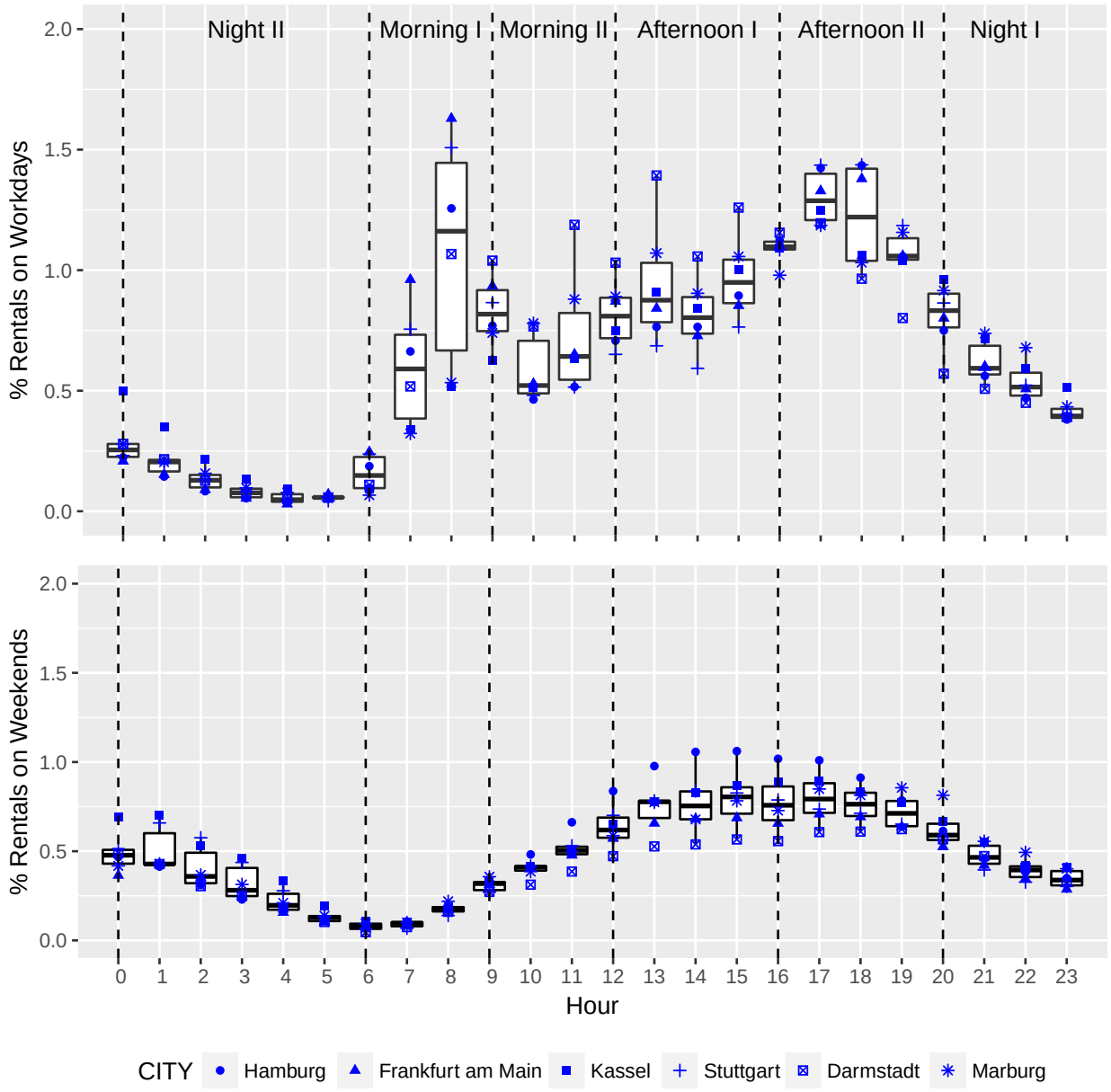


FIGURE 2: Hourly distribution and definition of times intervals

and a testing set of one city's zone. Validation was performed on a city level (and not using a random sample of zones) in order to examine how well the models would perform in a German city without a bike-sharing system. The city of Kassel was chosen for validation due to its high ridership and the fact that Hamburg and Frankfurt were not considered because they involved together around 76% of the zones of influence.

Stepwise OLS was considered in both directions, while BIC was chosen as a selection criterion. For GLM with lasso models, a Gaussian distribution was considered because it fit better the training data. A k–folds cross–validation (*35*) was implemented to calculate the shrinkage factor that helped the models to fit better the data. Concerning GBM, K–fold cross–validation was

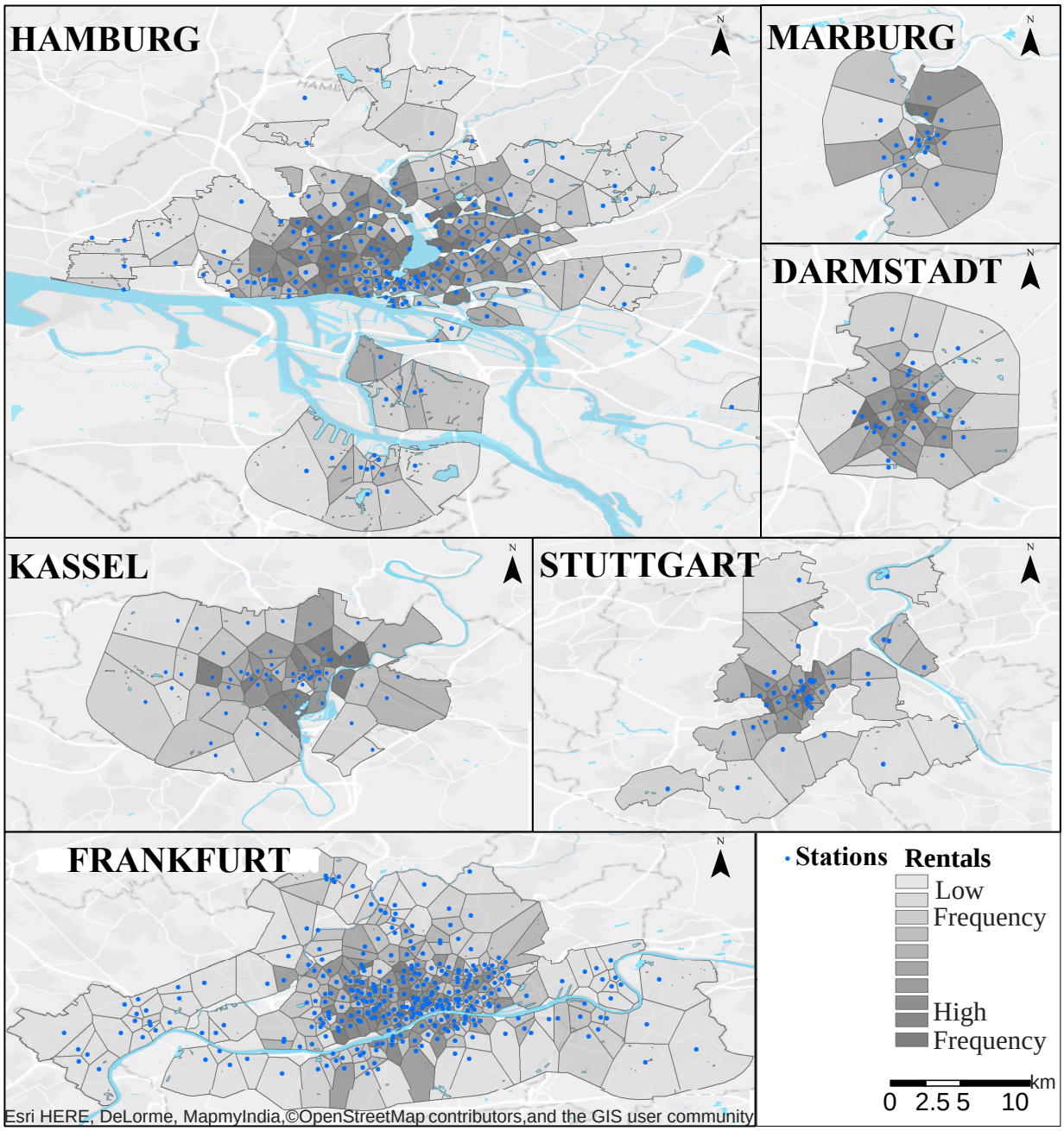


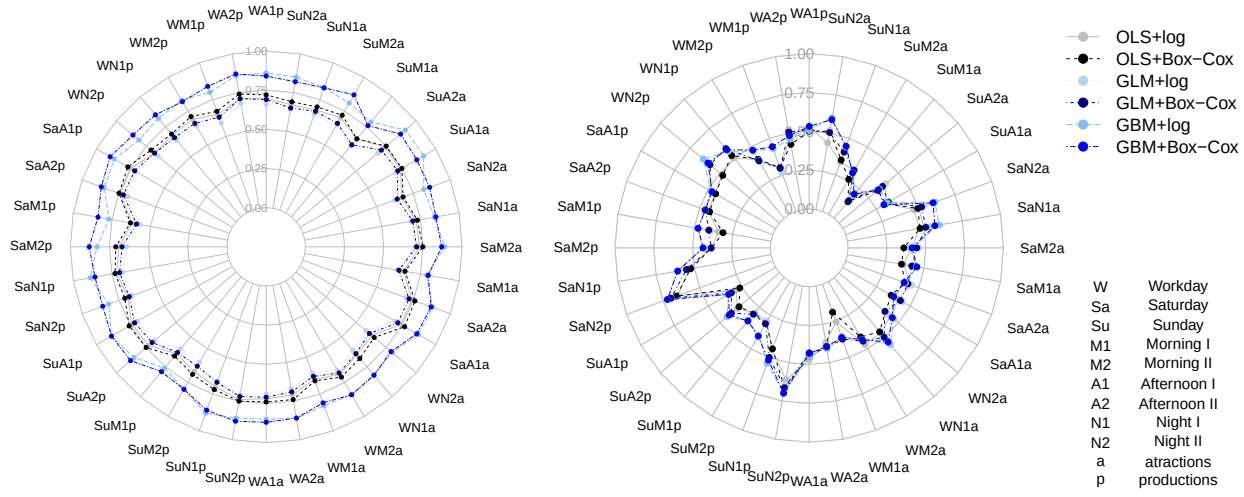
FIGURE 3: Spatial rentals distribution in cities of the study

realized to find the better number of trees or iterations with an input of 5 folds, a shrinkage factor of 0.0001, and an interaction depth of 6. The presence of heteroscedasticity and nonnormality in stepwise OLS and GLM led to the selection of logarithmic and Box-Cox transformations. Although these properties were not identified using GBM, the transformations were also carried out for matters of completeness. Outliers analysis was performed that indicated that zones with zero arrivals and departures should be removed.

In total 324 models were built, considering arrivals and departures for three cases (workday,

Saturday, Sunday), 6 time intervals (morning, afternoon and night at peak and off–peak periods), 3 regression modeling techniques (stepwise OLS, GLM, GBM) and 3 transformation techniques (no transformation, logarithmic and Box-Cox).

Regarding the parsimony the models, stepwise OLS selected the fewest number of variables with an average in all temporal scales of 15.55 variables, followed by GLM with 26.13 and finally, GBM with 39.90. According to the fitting results in the training cities (Figure 4a), R^2adj in the three regression methods trend together over different time periods. This indicates a rather indifference to time goodness-of-fit. Between the regression techniques, GBM usually presented higher R^2adj values. According to the validation performed with the city of Kassel, (Figure 4b) shows a slight difference between the R^2 values of different regression techniques, but there was a significant difference according to the time. Afternoons and nights showed the highest performances, especially during the weekends. Finally, in all cases, a logarithmic and a Box-Cox transformation illustrated a better goodness-of-fit with the logarithmic transformation to be slightly better.



(a) R^2 adjusted from fit models

(b) R^2 for validation case

FIGURE 4: Comparison of the R^2 adjusted from the fit values of different models

Based on the above results, we also performed cross-validation, by dividing ridership data in a training set of 5 cities' zones and a test set of one city's zone, for the case of GBM with a logarithmic transformation. Hamburg and Frankfurt were excluded from this analysis since they represent the majority of the zones of influence. The results presented in Figure 5 illustrate a rather high performance in most cases, with workdays to have less variation of the validation scores than on weekends. The city of Stuttgart, as a testing set, was the only case that showed a better performance than the city of Kassel.

Factors Affecting Bike Sharing

Aiming at constructing an overview of the factors found to affect ridership, the occurrence of parameters in all model structures were used. Figure 6 presents the most often selected variables by the regression techniques per time interval. Darker blue indicates higher selection frequency.

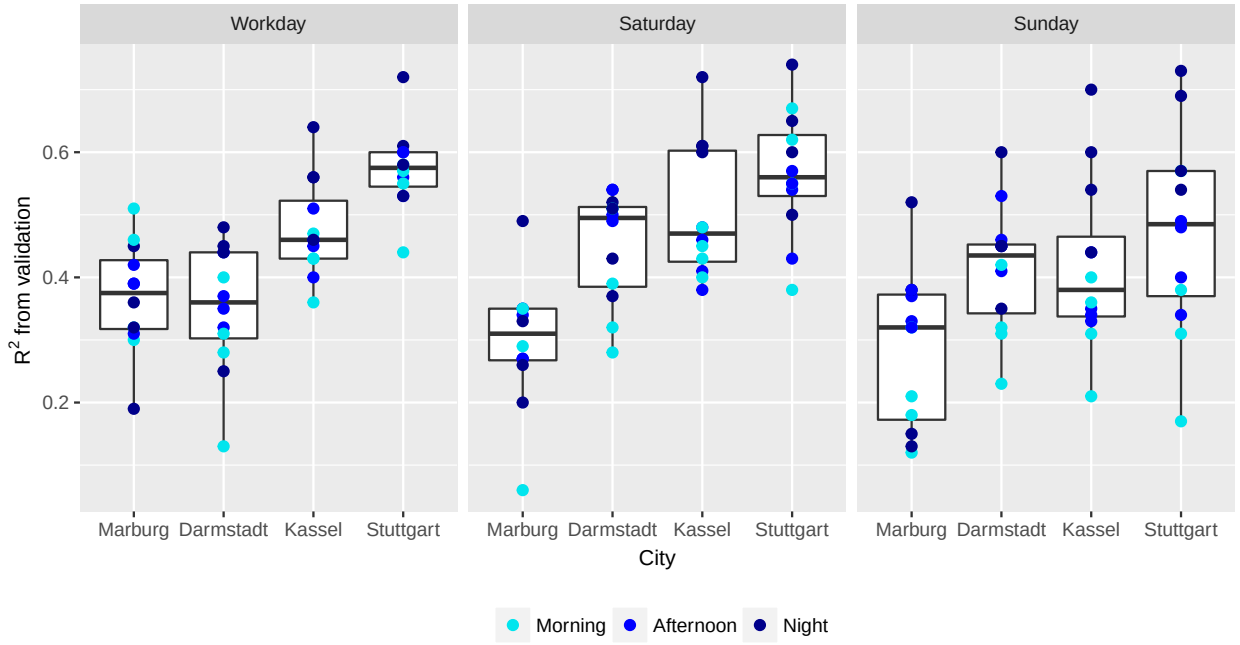


FIGURE 5: R^2 from validation by testing other cities (GBM with log transformation)

The most repetitive variables are the city population, the distance to the city center, bakeries, bicycle parking, memorials, residential areas and car sharing stations.

Multiple international cities approach

The international application focused on the SBBS systems *"Call a Bike"* in Hamburg (`www.callabike-interaktiv.de/de/staedte/hamburg`), *"Divvy"* in Chicago (`www.divvybikes.com`) and *"Bixi"* in Montreal (`montreal.bixi.com`). The main objectives of this application were the exploration of model transferability and the extraction of conclusions for the application of the methods for different city structures on an international level. These three cities where chosen since they share common characteristics as representative cities in their countries with border limited by a body of water. However, in terms of population, Montreal and Hamburg have relatively same inhabitants, but Chicago has around one million more inhabitants. Thus, we are referring to mainly large cities, with a rather high population that could have different travel characteristics and ridership patterns. As a consequence, the analysis was performed from the beginning guiding a somewhat different modeling and validation approach. Bixi-Montreal data (*36*) were collected from April 2014 until November 2017 with a data size of 734 MB and 545 stations. Divvy system works with 585 stations, where 1.75 GB data (*37*) was collected from June 2013 until December 2017. Finally, 2.5 GB of Call a bike rentals were collected in Hamburg from April 2014 until May 2017 in 207 stations (*38*).

Data analysis, and processing

The approach followed for the data analysis and processing was the same as the one for the national case, with the exception that the rentals data were aggregated at an additional seasonal level. The seasonality was added to analyze its effect on the resulting models. Chicago presented 9.93 rentals

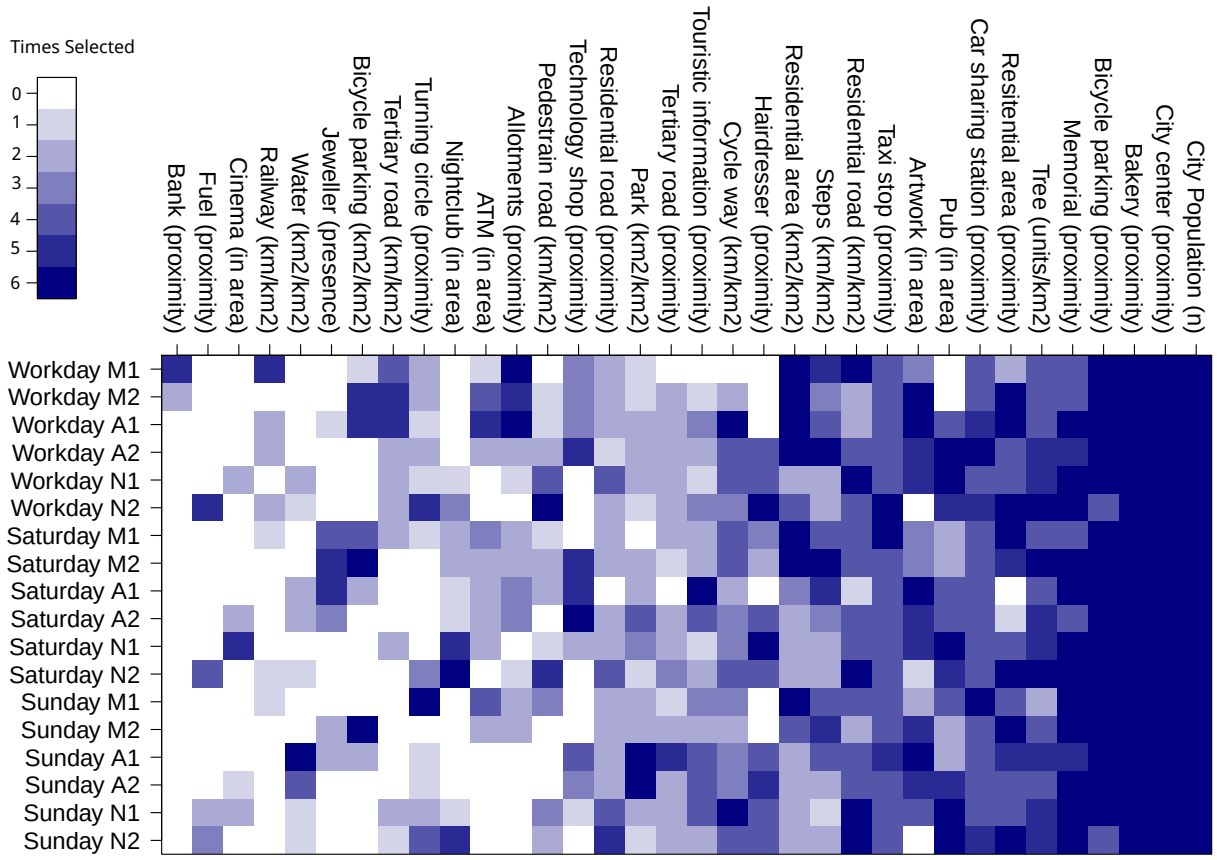


FIGURE 6: Repetitive outcome variables influencing the arrivals using the modeling techniques after logarithmic and Box-Cox transformations

per day interval per station, while Hamburg 24.59 and Montreal 16.47. Figure 7 shows the distribution of rentals per day interval in Chicago, Hamburg, and Montreal. These three cities present a relatively similar distribution with an exception on the day interval "Morning I" on weekends. Figure 8 illustrates some examples of the spatial distribution of the rentals per time interval. It shows higher ridership close to bodies of water. Concerning independent variables, **154** built environment variables were present in the three examined cities following the procedure as in the national approach, where finally **113** non-collinear variables were considered for the modeling procedure.

Model building, validation and variables selection

Stepwise OLS with a logarithmic transformation was used for the model building. The choice of using only one method was based on the computational time required to estimate all models and because in the national application it was the most parsimonious method while preserving relatively good fitting results. 72 models were built (one for each of four seasons, six day intervals, and workdays, Saturdays, and Sundays). On validation, we considered an alternative approach of the national case study by training 70% of the stations, and 30% for validation, without taking into account the cities boundaries.

Model fitting and validation scores for Hamburg, Chicago an Montreal are shown in Figure 9. R^2adj resulted of 0.68 as an average in the 72 models, and 0.63 as a R^2 score in the testing

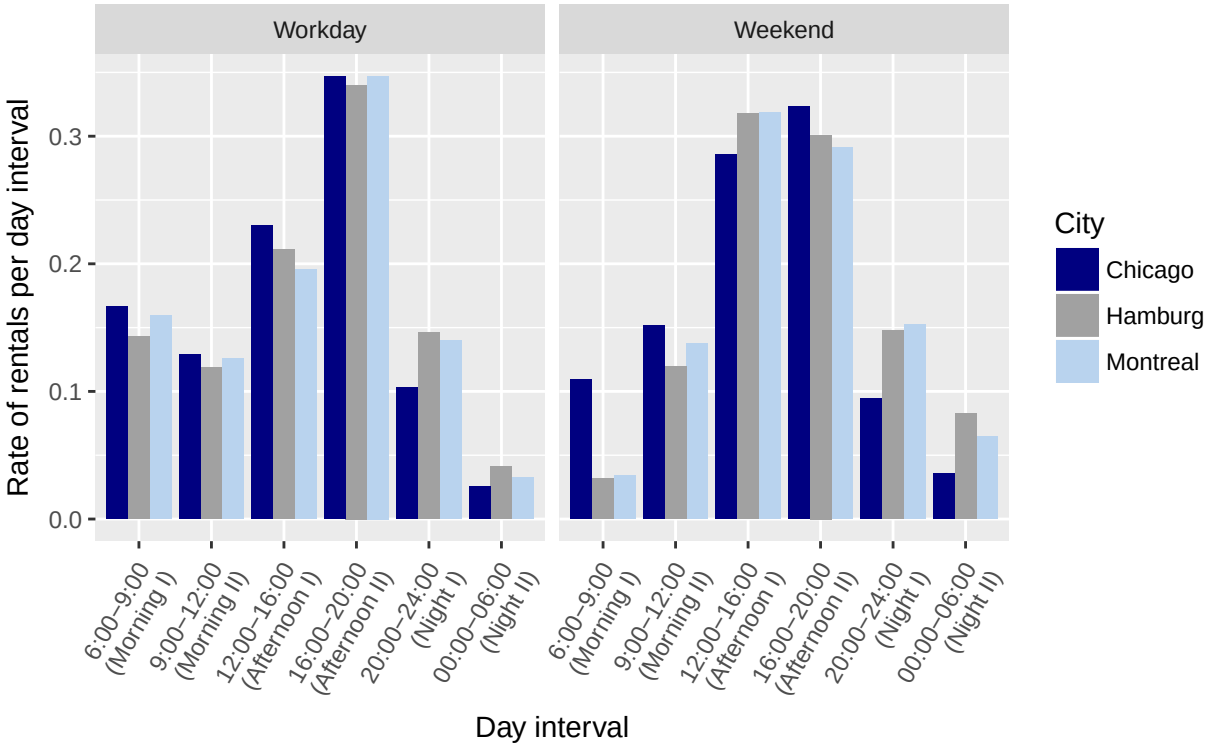


FIGURE 7: Distribution of rentals per day interval in Chicago, Hamburg and Montreal

process. We run five times the model building as a cross-validation process, where the R^2 from the validation varied on average around the third decimal. The lower validation results were during mornings on the weekends during all seasons, while higher values were associated with summer and winter.

As an example of the resulting influencing factors in summer and winter on workdays, Figure 10 presents the t-scores of the resulting models. On average 18.5 variables where selected per model. The most common selected variables were the population and the proximity to colleges and marina (area serving of leisure boats and yachts), bus stations, and restaurants and cafes. Mainly in summer, several variables of high significance which have a negative influence during morning and afternoon have a positive influence during the night. Land use influencing ridership was mainly residential and parks.

DISCUSSION

A data-driven method using exclusively open-source data was applied in two case studies considering multiple cities in a national and an international level. From around 800 possible built environment variables, the 144 most relevant and non-collinear variables were selected for the model building for the national approach, while 113 were selected for the international approach.

Concerning model applicability, linear and non-linear modeling techniques were tested in the national approach. GBM was the regression method that best fit the data, followed by GLM. GLM and GBM required cross–validation tests to select the input arguments that helped to build models. However, stepwise OLS was parsimonious with fewer input arguments, and its results were easier to interpret. The three modeling techniques presented similar validation

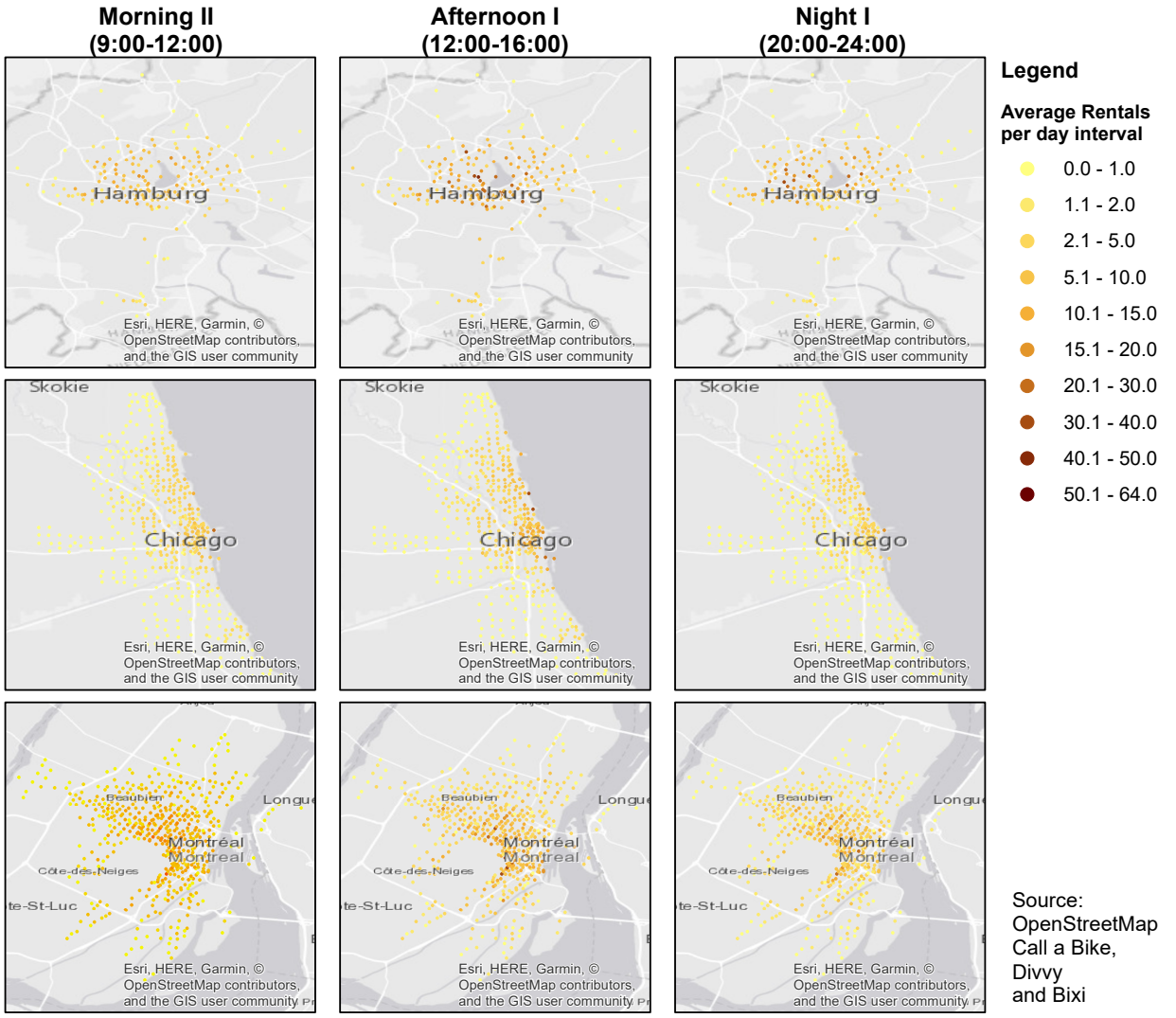


FIGURE 8: Spatial distribution of rentals per day interval in Chicago, Hamburg and Montreal

results. Logarithmic and Box-Cox transformations helped the models to predict better the arrivals and departures and to select logical variables that would influence the shared bicycles ridership. Generally, for the three regression methods, the logarithmic transformation performed a higher R^2 in the validation phase.

The advantage of stepwise OLS and GLM was that a variable selection process was implicit in the methods, but for GBM a variable selection process had to be developed to select those with more influence from the ranking list. The most influencing 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 large like Hamburg or small like Marburg. The distance to the city center played a significant role for ridership as seen in Figure 2. The third most influencing variable is the distance to bakeries. If a station is close to a bakery, this increases the probability of higher ridership at that station.

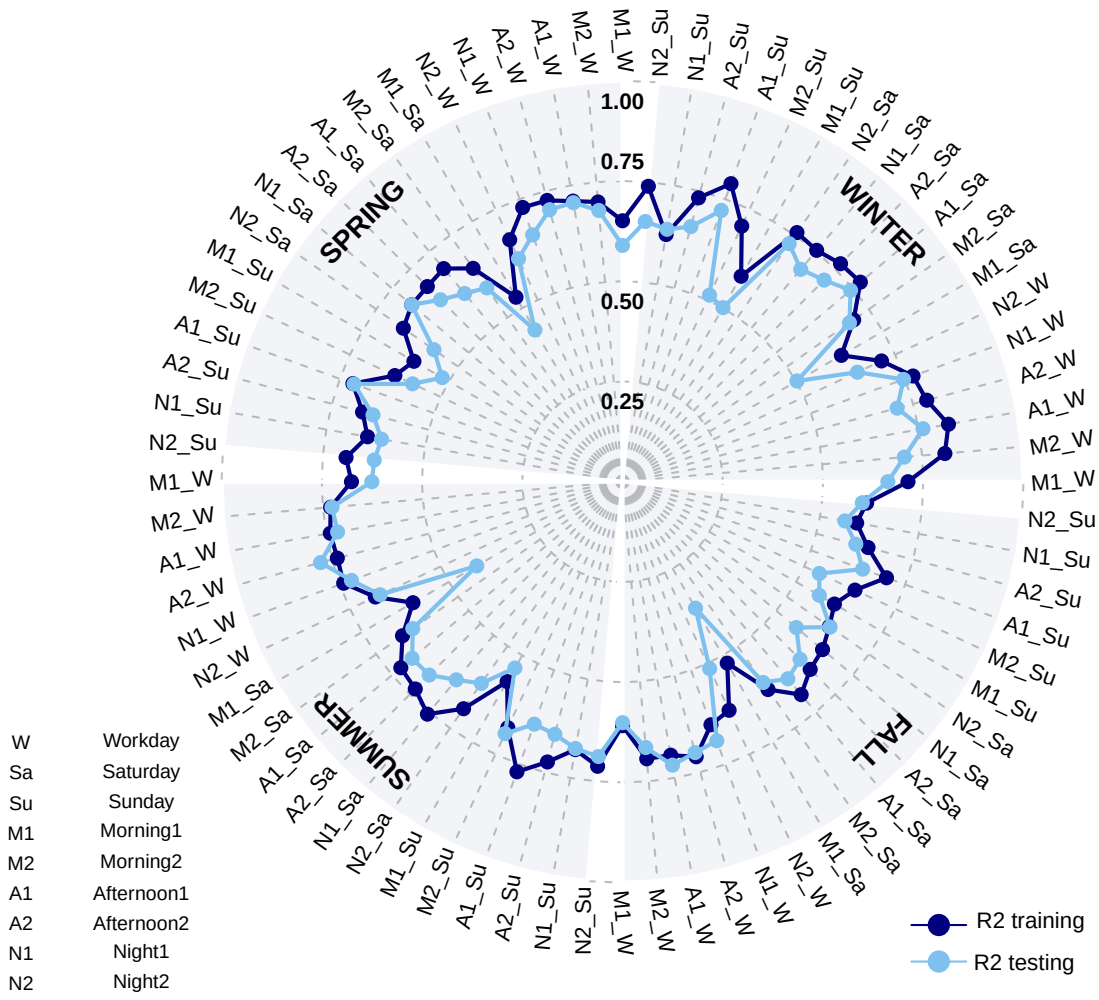


FIGURE 9: Model fitting and validation scores for Hamburg, Chicago and Montreal (70% of the total stations for training)

In all developed models, several selected variables were logically correlated to bike sharing ridership, were similar to literature review findings (Table 1), and they were coherent on the authors' expectations in influencing the arrivals and departures of bike sharing. For instance, the most influencing variables are related to leisure activities, parks, green areas, and bodies of water on the weekends, banks in the morning, gas stations, pubs, cinemas, clubs at night, shops on Saturdays, and memorials outside of working hours. Just a few transport–related variables significantly influenced the models. Distance to a car sharing station was significant for all time intervals as well as bicycle parking. The only public transport variable displayed was railways during workday mornings. It is worth mentioning that the tram and metro variables were not considered, because they were not present in most of the studied cities.

According to the international approach, stepwise OLS with a logarithmic transformation was chosen after the benefits identified in the national approach. On average, 18.5 variables were selected per model. Urban structure was found to play an important role based on the distance from all stations to the marina and colleges and also land use represented by residential area and

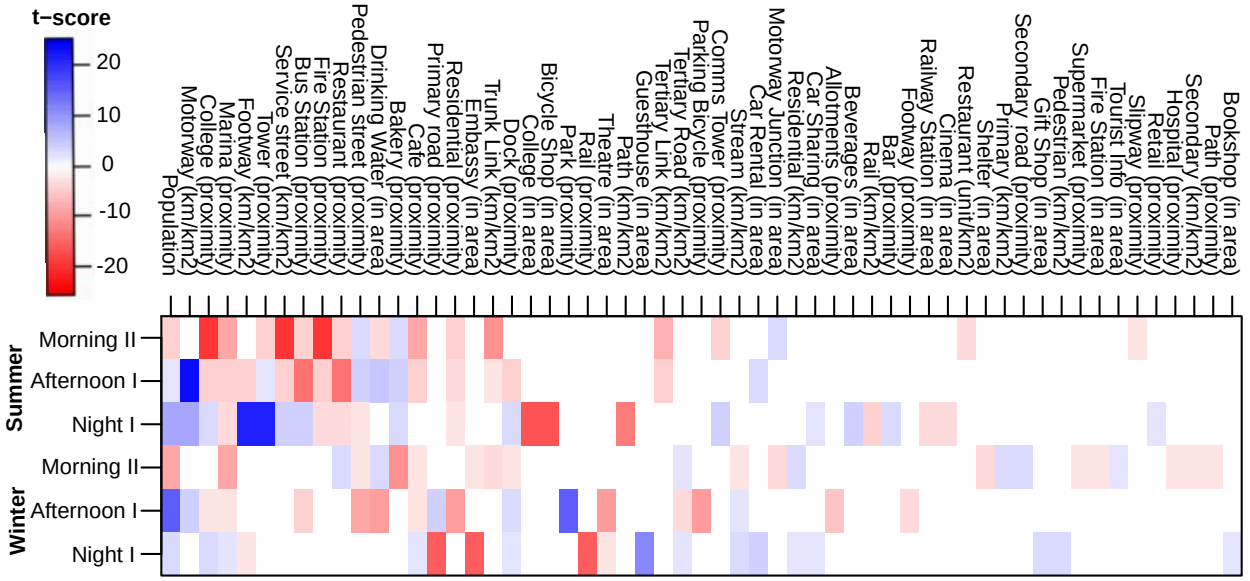


FIGURE 10: Example of t-scores for the international approach (Workdays)

parks. Summer and winter presented different factors, for example, bakeries, bus stations, and restaurants are more significant in summer and not in winter. Logical variables were present as the influence of bar and railway station in summer night or colleges influencing negatively at night. An important observation is a correlation with car sharing stations during the night representing a possible correlation between car and bike sharing.

Both approaches showed that the modeling validation results were correlated to the hourly ridership distribution. Similar relative ridership hourly distribution was associated with higher scores. For example, in the international approach in morning on weekends showed the most different distribution between the cities (Figure 7), presenting the worst modeling performance. On the other hand, in the national approach, the models that fit better the data were in the afternoon and at night where the different cities showed a less variance in the bike sharing usage (Figure 3). At these times, models presented better results from the validation and illustrated a more logical selection of variables that influenced ridership. Also, on weekends the rate of ridership distribution was more similar between the cities than on the workdays (Figure 3) showing higher validation results. Finally, it was found that the modeling results did not depend on the size of the cities but on the similarity of the distribution of the rate of rentals.

CONCLUSION

To the best of the authors' knowledge, this is the first study that analyzes factors affecting bike sharing systems ridership on a local level in multiple cities. The resulting influencing factors are not only based on one city but beyond the geographic boundaries, which will help to use the resulting models to forecast the bike sharing usage in a different city. A data–driven method was developed to analyze the influence of the built environment in the rentals of station–based bike sharing systems in multiple cities. An original approach was considered by modeling different cities with different sizes in two case studies 1) on a national level and 2) on an international level.

GBM with a logarithmic transformation of the dependent variables were found to vali-

date slightly better the dataset. Stepwise OLS and a logarithmic transformation of the dependent variables was found to select fewer variables than other models without decreasing the validation results significantly. In Germany, 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 historical bike sharing rentals over time intervals were displayed, such as higher arrivals on nights close to pubs, cinemas, and nightclubs; or the presence of bodies of water, parks or green areas on Sundays. On an international level, the distance to the marina and colleges played an important role. Different influencing factors were present between different seasons.

The data–driven method will help as a decision-making support tool to implement or expand bike sharing systems. Transport planners will have in a relatively brief time an idea of hot and cold spots of arrivals and departures of bike sharing systems to help them set coverage areas and place stations. This method can also show the validity and increase the reliability of measures, policies, and shared mobility projects. Further research and implementation can be developed as improvements and expansion of the case of study (including more variables, such as topography or population density) or correlations between factors influencing car and bike sharing.

ACKNOWLEDGMENT

This study was partially supported by “Hans Boeckler Stiftung” and the MO3 project granted by the International Graduate School of Science and Engineering of TUM. The authors also are thankful for the bike sharing systems Divvy Bikes, BIXI–Montreal and "Call a bike" for sharing the rentals data through an open portal.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm the contribution to the paper as follows: study conception and design: D. Duran, E. Chaniotakis, C. Antoniou; data collection: D. Duran; analysis and interpretation of results: D. Duran, E. Chaniotakis, C. Antoniou; draft manuscript preparation: D. Duran, E. Chaniotakis, C. Antoniou. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

[1] Büttner, J. and T. Petersen, *Optimising Bike Sharing in European Cities: A Handbook*. OBIS, 2011.

[2] Böckmann, M., 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*, 2013.

[3] Shaheen, S., E. Martin, A. Cohen, and R. Finson, *Public Bikesharing in North America: Early Operator and User Understanding, MTI Report 11-19*. Mineta Transportation Institute, 2012.

[4] Meddin, R. and P. DeMaio, *The bike-sharing world map*. http://www. metrobike. net, Accessed on 21.06.2018, 2018.

[5] Firnkorn, J. and S. Shaheen, 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*, Vol. 113, 2016, pp. 897 – 909.

[6] Shaheen, S., S. Guzman, and H. Zhang, Bikesharing in Europe, the Americas, and Asia: past,

present, and future. *Transportation Research Record: Journal of the Transportation Research Board*, , No. 2143, 2010, pp. 159–167.

[7] DeMaio, P., Bike-sharing: History, impacts, models of provision, and future. *Journal of public transportation*, Vol. 12, No. 4, 2009, p. 3.

[8] Hamann, T. K. and S. Güldenberg, Overshare and Collapse: How Sustainable are Profit-Oriented Company-to-Peer Bike-Sharing Systems?, 2017.

[9] Nikitas, A., *Bike-sharing fiascoes and how to avoid them – an expert's guide*. https://theconversation.com/bike-sharing-fiascoes-and-how-to-avoid-them-an-experts-guide-84926 , 2017, accessed on: 17.07.2018.

[10] Chardon, C. M. D., G. Caruso, and I. Thomas, Bicycle sharing system 'success' determinants. *Transportation Research Part A: Policy and Practice*, Vol. 100, 2017, pp. 202 – 214.

[11] Zhao, J., W. Deng, and Y. Song, 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*, Vol. 35, 2014, pp. 253 – 264.

[12] Faghih-Imani, A. and N. Eluru, Incorporating the impact of spatio-temporal interactions on bicycle sharing system demand: A case study of New York CitiBike system. *Journal of Transport Geography*, Vol. 54, 2016, pp. 218 – 227.

[13] El-Assi, W., M. Mahmoud, and K. Habib, Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. *Transportation*, Vol. 44, No. 3, 2017, pp. 589–613.

[14] Tran, T. D., N. Ovtracht, and B. F. d'Arcier, Modeling Bike Sharing System using Built Environment Factors. *Procedia CIRP*, Vol. 30, 2015, pp. 293 – 298, 7th Industrial Product-Service Systems Conference - PSS, industry transformation for sustainability and business.

[15] Faghih-Imani, A., R. Hampshire, L. Marla, and N. Eluru, An empirical analysis of bike sharing usage and rebalancing: Evidence from Barcelona and Seville. *Transportation Research Part A: Policy and Practice*, Vol. 97, No. Supplement C, 2017, pp. 177 – 191.

[16] Noland, R. B., M. J. Smart, and Z. Guo, Bikeshare trip generation in New York City. *Transportation Research Part A: Policy and Practice*, Vol. 94, 2016, pp. 164 – 181.

[17] Wang, X., G. Lindsey, J. Schoner, and A. Harrison, Modeling bike share station activity: Effects of nearby businesses and jobs on trips to and from stations. *Journal of Urban Planning and Development*, Vol. 142, No. 1, 2015, p. 04015001.

[18] Faghih-Imani, A., N. Eluru, A. M. El-Geneidy, M. Rabbat, and U. Haq, How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal. *Journal of Transport Geography*, Vol. 41, 2014, pp. 306 – 314.

[19] Mattson, J. and R. Godavarthy, Bike share in Fargo, North Dakota: Keys to success and factors affecting ridership. *Sustainable Cities and Society*, Vol. 34, 2017, pp. 174 – 182.

[20] Fishman, E., S. Washington, and N. Haworth, Bike Share: A Synthesis of the Literature. *Transport Reviews*, Vol. 33, No. 2, 2013, pp. 148–165, cited By 98.

[21] Schmöller, S. and K. Bogenberger, Analyzing External Factors on the Spatial and Temporal Demand of Car Sharing Systems. *Procedia - Social and Behavioral Sciences*, Vol. 111, 2014, pp. 8 – 17.

[22] Bishara, A. and J. Hittner, Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological methods*, Vol. 17, No. 3, 2012, p. 399.

1 [23] Box, G. E. P. and D. R. Cox, An Analysis of Transformations. *Journal of the Royal Statistical*
2 *Society. Series B (Methodological)*, Vol. 26, No. 2, 1964, pp. 211–252.
3 [24] Chatterjee, S. and A. Hadi, *Regression analysis by example*. John Wiley and Sons, 2015.
4 [25] Tibshirani, R., Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Sta-*
5 *tistical Society. Series B (Methodological)*, Vol. 58, No. 1, 1996, pp. 267–288.
6 [26] Friedman, J., Greedy function approximation: A gradient boosting machine. *Annals of Statis-*
7 *tics*, Vol. 29, No. 5, 2001, pp. 1189–1232, cited By 2353.
8 [27] Lin, D., D. P. Foster, and L. H. Ungar, VIF Regression: A Fast Regression Algorithm for
9 Large Data. *Journal of the American Statistical Association*, Vol. 106, No. 493, 2011, pp.
10 232–247.
11 [28] Akaike, H., Maximum likelihood identification of Gaussian autoregressive moving average
12 models. *Biometrika*, Vol. 60, No. 2, 1973, pp. 255–265.
13 [29] Schwarz, G., Estimating the dimension of a model. *The annals of statistics*, Vol. 6, No. 2,
14 1978, pp. 461–464.
15 [30] Friedman, J., T. Hastie, and R. Tibshirani, *The elements of statistical learning*, Vol. 1.
16 Springer series in statistics New York, 2001.
17 [31] Willing, C., K. Klemmer, T. Brandt, and D. Neumann, Moving in time and space – Location
18 intelligence for carsharing decision support. *Decision Support Systems*, Vol. 99, 2017, pp. 75
19 – 85, location Analytics and Decision Support.
20 [32] Deutsche Bahn AG, *Das smarte Leihfahrrad der Deutschen Bahn | Call a Bike*.
21 https://www.callabike-interaktiv.de/de, 2017, accessed on: 31.10.2017.
22 [33] OpenStreetMap-contributors, *Planet dump retrieved from https://planet.osm.org* .
23 https://www.openstreetmap.org , 2017, accessed on: 31.10.2017.
24 [34] Statistisches Bundesamt, *Bevölkerung und Erwerbstätigkeit. Bevölkerung mit Migrationsh-*
25 *intergrund. Ergebnisse des Mikrozensus 2011. Fachserie 1, Reihe 2.2*, 2012, accessed on:
26 31.10.2017.
27 [35] James, G., D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*,
28 Vol. 112. Springer, 2013.
29 [36] BIXI-MONTREAL, *Open Data - BIXI Montréal*. www.bixi.com/en/open-data, 2017, ac-
30 cessed on 17.12.2017.
31 [37] Bikes, D., *Divvy Data*. https://www.divvybikes.com/system-data, 2018, accessed on:
32 15.01.2018.
33 [38] Deutsche Bahn (DB), *Call A Bike - Open-Data-Portal – Deutsche Bahn Datenportal*.
34 http://data.deutschebahn.com/dataset/data-call-a-bike, 2017, accessed on: 31.10.2017.