geographical analysis

Check for updates

Geographical Analysis (2023) **55**, 651–684

Prediction of Bike-sharing Trip Counts: Comparing Parametric Spatial Regression Models to a Geographically Weighted XGBoost Algorithm

Katja Schimohr[1], Philipp Doebler[2], and Joachim Scheiner[1]

[1]Department of Spatial Planning, Transport Research Group, Technische Universität Dortmund, Dortmund, Germany, [2]Department of Statistics, Research Group of Statistical Methods in Social Sciences, Technische Universität Dortmund, Dortmund, Germany

Regression models are commonly applied in the analysis of transportation data. This research aims at broadening the range of methods used for this task by modeling the spatial distribution of bike-sharing trips in Cologne, Germany, applying both parametric regression models and a modified machine learning approach while incorporating measures to account for spatial autocorrelation. Independent variables included in the models consist of land use types, elements of the transport system and sociodemographic characteristics. Out of several regression models with different underlying distributions, a Tweedie generalized additive model is chosen by its values for AIC, RMSE, and sMAPE to be compared to an XGBoost model. To consider spatial relationships, spatial splines are included in the Tweedie model, while the estimations of the XGBoost model are modified using a geographically weighted regression. Both methods entail certain advantages: while XGBoost leads to far better values regarding RMSE and sMAPE and therefore to a better model fit, the Tweedie model allows an easier interpretation of the influence of the independent variables including spatial effects.

Introduction

Analysis of transportation data gives insight into the functioning of urban transport systems. A large number of emerging shared mobility options including car-, bike-, scooter- and e-scooter-sharing, taxis and ride-hailing services such as Uber generate usage data including origin and destination of trips. In this research, we focus on shared micromobility services. Analyzing starting and ending points of such trips allows to efficiently plan and manage sharing systems (Zhao et al. 2019; Reck et al. 2021). In urban planning, origin-destination data can be used to evaluate the usage of sharing systems in order to learn about their functionality or to create regulations and measures that ensure an efficient integration into the urban transport

Correspondence: Katja Schimohr, Department of Spatial Planning, Transport Research Group, Technische Universität Dortmund, 44227 Dortmund, Germany.
e-mail: katja.schimohr@tu-dortmund.de

Submitted: February 8, 2022. Revised version accepted: October 28, 2022.

doi: 10.1111/gean.12354

system (Sun, Li, and Zuo 2019; Caspi, Smart, and Noland 2020; Cheng et al. 2020). The analysis of existing bike-sharing systems can also help cities to estimate usage and locate stations or usage areas before implementing a new system, based on the spatial features of a city (Lee and Sener 2020).

Different statistical methods can be employed for this task. While traditionally, parametric methods such as linear or negative binomial (NB) regression are used to model this kind of data, machine learning methods have emerged that could potentially outperform regression models in terms of predictive accuracy (Ramesh et al. 2021). Accuracy benefits also because of a decrease in biases occurring when neglecting spatial structure (Dormann et al. 2007; Meyer et al. 2019), potentially improving planning decisions. Models dealing with the spatial distribution of trip data face three main challenges: they have to account for spatial autocorrelation and be able to handle count data as well as overdispersion. So far, there is little research comparing parametric and machine learning methods for the analysis and prediction of spatial bike-sharing data that are able to meet these three requirements. Our study is novel in doing exactly this.

This article aims at comparing the predictive abilities of a parametric regression model to those of a machine learning approach using a geographically weighted eXtreme Gradient Boosting (XGBoost) algorithm. While a parametric regression model offers the advantage of easier interpretability, a machine learning model may generate better predictions provided that appropriate features are used for fitting (Sathishkumar, Park, and Cho 2020; Ramesh et al. 2021). Therefore, the objective of this research is to compare both methods in the context of modeling spatial count data. Among other characteristics, such as computation time and complexity of model definition, especially the accuracy of predictions and ease of interpretation are to be evaluated for both approaches.

In this article, the bike-sharing system of *nextbike* in Cologne, Germany, is examined as a case study. The bikes are organized free-floating, allowing for flexible rental and drop-off throughout the operation area. Usage data is freely available and was collected over a period of five weeks in 2019. The number of bike-sharing trips that started within 100×100 m grid cells is modeled as a function of various spatial factors.

While most studies applying machine learning methods on bike-sharing data deal with the temporal modeling and forecasting of usage demand, the focus of this research lies on spatial factors. Temporal influence factors on bike-sharing such as weekday, time of day or weather-related factors have been extensively analyzed in other studies applying both parametric and machine learning methods (see e.g., Gebhart and Noland 2014; Kaviti et al. 2018; Gong and Yamamoto 2019; Sathishkumar, Park, and Cho 2020; Hu et al. 2021), while the spatial influence factors are of interest from a geographical perspective. This is why spatial models and their characteristics are examined in this article.

Literature review

The spatial data generated by car-/bike-/scooter-sharing systems or ride-hailing systems has been analyzed through the application of various statistical methods that are presented in the following. Liao and Correia (2020) present a literature review of papers that deal with electric carsharing, e-bike- and e-scooter-sharing including a few that analyze spatial data sets.

Parametric models

Linear regression is used to investigate spatial influence factors on bike-sharing by Wang et al. (2015) and Tran and Ovtracht (2018). Due to the distribution of the dependent variable, the

number of trips per station or area, usually Poisson or NB regression has been found to be a more adequate model for this task (Gong and Yamamoto 2019; Huo et al. 2021). Poisson regression is also used to model other crowd-sourced bicycling data (Sanders et al. 2017; Roy et al. 2019). Bai and Jiao (2020) analyze the influence of various spatial factors on scooter-sharing using NB regression models. To account for an excess number of zeros, zero-inflated regression is used to examine the influence of various spatial factors on car sharing in Berlin (Wagner, Brandt, and Neumann 2016), zero-inflated NB regression is used in the analysis of bike-sharing reallocation (Zhao et al. 2019) or other transport-related questions as well (Liu et al. 2018a).

To deal with the issue of overdispersion in the modeling of spatial count data, advanced models have been developed, for example by Mardalena et al. (2022) and Cepeda-Cuervo, Córdoba, and Núñez-Antón (2018). Nevertheless, the choice of distribution a model is based on makes a difference as well. Da Silva and Rodrigues (2013) develop a geographically weighted NB regression model and compare it to a NB and a geographically weighted Poisson regression, proving its advantage in reducing overdispersion. Chen et al. (2020) confirm that a geographically weighted NB model can account for overdispersion and outperforms a model based on a Poisson distribution as well.

A large number of studies meets the issue of spatial autocorrelation through the estimation of a Geographically Weighted Regression (GWR): Gong and Yamamoto (2019) apply a spatially filtered NB regression model in the analysis of bike-sharing. Bao, Shi, and Zhang (2018) apply *k*-means clustering on bike-sharing stations and then analyze the impact of spatial influence factors using a GWR. Ji et al. (2018) fit a Poisson GWR to examine the influence of various spatial factors on bike-sharing. Caspi, Smart, and Noland (2020) apply a GWR to analyze the influence of spatial factors on e-scooter trip destinations. Perlman and Roy (2021) analyze spatial influence factors on trips taken by Uber using a GWR. Wang et al. (2020b) apply a GWR on bike-sharing data .

Other methods to account for spatial autocorrelation include the estimation of spatial error and spatial lag models as applied by Correa, Xie, and Ozbay (2017) to compare taxi and Uber demand. Furthermore, Wang et al. (2021b) estimate a linear regression model with graph regularization. Hu et al. (2021) account for spatial autocorrelation in the analysis of bike-sharing data through the inclusion of an additive term in the Generalized Additive Model (GAM) expressing the spatial coordinate interaction.

Several papers compare the qualities of regression methods. Sawalha and Sayed (2006) compare Poisson and NB regression. Wang et al. (2015) affirm the advantages of NB regression over log-linear ordinary least squares (OLS) regression regarding bike-sharing data. Zhang, Cheng, and Jin (2019) compare geographically weighted OLS regression to geographically weighted NB regression in the analysis of an origin-destination data set of traffic flows and conclude that the NB regression performs better. Hosseinzadeh et al. (2021) compare an OLS regression to a GWR in modeling e-scooter trips in relation to spatial influence factors. Yang et al. (2020a) compare an OLS, a GWR and a semi-parametric GWR model to analyze spatial influence factors on bike-sharing. Here, the semi-parametric GWR outperforms the other two models.

Machine learning approaches

Machine learning approaches are increasingly applied to model and predict usage of sharing systems, even though most studies model the temporal distribution to predict demand over time. Zhang et al. (2021) analyze the specific routes that are undertaken by e-scooter using a recursive logit route choice model. Cheng et al. (2021) analyze spatial factors on car sharing demand combining machine learning and generalized linear models and explanations based on

SHAP values. Wang et al. (2020a) apply a long short-term memory (LSTM) recurrent neural network in the temporal modeling of car sharing demand of different stations. Wang, Hu, and Jiang (2021a) apply gradient boosting regression trees for the prediction of car sharing usage. XGBoost has already been applied in the context of sharing systems as well (Sathishkumar, Park, and Cho 2020; Yang et al. 2020b; Ramesh et al. 2021). A *Kaggle* competition dealt with the modeling of daily usage in the bike-sharing system of Washington D.C., with 3,242 teams entering the competition (Kaggle 2015). Here, the winning entry applies XGBoost to model the bike-sharing demand (Yang et al. 2020b).

Many studies do not only implement a model using XGBoost but compare the performance of several methods regarding computation time and modeling outcomes. Sometimes, XGBoost is also used as a competitive reference approach when new methods are developed (Liu, Shen, and Zhu 2018b). Ramesh et al. (2021) model bike-sharing demand at stations comparing random forest, extreme gradient boosting, and linear regression. Here, XGBoost performs best regarding goodness of fit which was measured using R^2. Sathishkumar, Park, and Cho (2020) compare linear regression, a gradient boosting machine, a support vector machine, boosted trees, and XGBoost to model the temporal distribution of bike-share rentals. In this analysis, the gradient boosting machine leads to the best results considering R^2, root mean square error (RMSE), Mean Absolute Error (MAE) and the Coefficient of Variation (CV), closely followed by XGBoost. Yang et al. (2020b) apply a graph-based approach while also incorporating XGBoost in their study on bike-sharing. Li and Axhausen (2019) compare a whole range of different models including time-series models, machine learning approaches such as random forest or XGBoost, and deep learning approaches such as LSTM recurrent neural network to model taxi demand for two different data sets. The authors cannot determine one single best model. Instead, different models perform best according to different evaluation metrics. XGBoost ranks among the best models regarding computation time and symmetric mean absolute percentage error (sMAPE), while an LSTM leads to slightly better RMSE values. Alencar et al. (2021) evaluate the qualities of LSTM for temporal modeling of car sharing demand to several other modeling approaches including XGBoost. Again, different methods perform best for different tasks and regarding different performance measures (MAE and RMSE).

In sum, a vast range of methods has already been applied in the spatial modeling of different sharing systems and there is no standard method. The choice of the best method depends on the specific research interest and data set. Temporal and spatiotemporal models have been the focus of recent work, but purely spatial methodology is most relevant for geographic analyses. This article aims at comparing two methods that meet the requirements of count data, an excess number of zeros and spatial autocorrelation simultaneously. Generalization and interpretation is often easier based on parametric regression models, but predictive performance might be lower compared to current machine learning methods. Based on the literature review, NB regression and GWRs stand out among the parametric regression models. XGBoost could prove to be a fast and accurate approach. Still, so far this method has been rarely applied in the modeling of spatial data. The following comparison will shed light on the trade-off between computational cost, predictive accuracy, and interpretability.

Data and methods

In the following, the generation and preprocessing of the data set used in this analysis are presented. For preprocessing, the geographic information system ArcGIS Pro is used in

combination with R, version 4.0.5 (R Core Team 2021), the analysis is executed in R. For data management, the packages `dplyr, lubridate`, for the preparation of figures and maps, additionally `RColorBrewer, lattice, sf`, and `cartography` are used.

Data

The study area for this research is the operating area of the *nextbike* bike-sharing system in Cologne, also called *flexzone*. Cologne is located in the West of Germany and inhabited by about 1.1 million residents, thus being the most populous city in the federal state North Rhine-Westphalia (Stadt Köln 2019, p. 11). The city covers an area of about 405 km^2 and is located on both sides of the river Rhine. The *nextbike flexzone* covers an area of about 84 km^2 in the central districts of Cologne on both sides of the Rhine. Bikes can be rented and returned flexibly throughout this area, but cannot be dropped off in parks nor on private property.

The data set to be modeled consists of the locations of all bikes in the *nextbike* system in Cologne during the observation period from September 30, 2019, 10:46 am until November 4, 2019, 10:49 am. The data were scraped from the website https://offenedaten-koeln.de using an excel VBA script that saved the locations every 15 min in a csv-table. All tables are combined in R and all changes in location indicating the completion of a trip are extracted. Using ArcGIS Pro, all spatial outliers, meaning trips that started or ended outside of Cologne, are removed. Additionally, trips shorter than 100 m are removed to exclude especially small position changes of parked bikes. The final data set contains 2,528,567 locations of bikes and 76,859 trips. The operating area of *nextbike* is divided into 8,955 grid cells of 100×100 m size. Based on this, the number of bike-sharing trips is aggregated spatially on the grid cell level and temporally over the whole length of the observation period. The dependent variable to be modeled is the total number of trips that started within each grid cell. Here, we find areas with low expected counts and even potential zero inflation, because there is little demand in the outer parts of the study area and it is not allowed to drop off bicycles within certain (parts of) grid cells due to the coverage by buildings, water or parks. In a small number of grid cells, there is a high demand and subsequently, there are very high trip numbers. This combination leads to overdispersion relative to the Poisson distribution.

In the resulting data set it is not possible to detect repositioning trips made by nextbike staff. As the share of repositioning trips among all trips in the system was found to range between 10% and 19% in several German station-based bike-sharing systems (Rabenstein 2015, p. 138), this should be kept in mind in the analysis of bike-sharing usage. Still, it should not have a strong influence on the results of the subsequent analysis evaluating different modeling approaches, as no grid cell should be disproportionally affected by this issue.

Additionally, it is possible that some trips could not be captured in our data set if a bike was dropped off and picked up again within less than 15 min. The duration between drop off and the next pick up of a shared bike has been defined as time to booking (ToB) (Guidon, Becker, and Axhausen 2019). This value strongly varies between bike-sharing systems, depending on the usage rate. In a free-floating e-bike-sharing system in Zürich, the average ToB amounts to 10.6 h, while in a bike-sharing system in Shanghai, ToB lies at 175.1 mins and in 35% of all times, bikes are rented again less than 10 min after drop off (Guidon, Becker, and Axhausen 2019; Li et al. 2020). Here, each bike is used 3.98 times a day on average, compared to 1.98 times in Cologne. The ToB lies at approximately 8.6 h. Therefore, we can expect only a small rate of missed trips. Since the proportion of missed trips can be assumed to grow with the

demand in a grid cell, especially in grid cells with a high demand individual trips might have been missed. Nevertheless, it should not strongly distort the measured trip numbers for certain grid cells.

For evaluation purposes, the data is split into a training and a test data set that contain approximately the same number of trips using the 3,263 different times of observation. All trips that started during the first time of observation are assigned to the training data set, trips that started during the second period 15 mins later are assigned to the test data set. Subsequently, all trips are assigned alternating with every time of observation, until the second last time, as it is not possible to observe new (completed) trips starting in the last time of observation. Therefore, each data frame contains the trips starting in 1,631 different times of observation. There are 38,875 trips included in the training data set and 37,965 trips included in the test data set. For both data sets, the number of trips that started through the whole observation period are summarized per grid cell. The training data set is used to train the models, while the test data set is used to generate predictions that can be compared to the observed values.

A range of possible spatial influence factors on bike-sharing is selected based on previous research (Tran and Ovtracht 2018; Yang et al. 2021). A complete list and data sources are included in Appendix S1. Different land uses might support bike-share usage or prevent to drop off bikes altogether (Hu et al. 2021). Therefore, the shares of different types of land use of each grid cell as defined in the Flächennutzungsplan = land use plan (FNP) of Cologne are included in the analysis. The share of green spaces and buildings is included as it is not allowed to drop off bikes in parks or on private property. Previous studies registered a higher usage in or near the central business district or in the vicinity of certain Points of Interest (POI) (Wang et al. 2015). Therefore, the number of shops, food outlets and bars is included as an indicator of centrality. All three types of POI are combined to avoid high multicollinearity of the original variables. Additionally, various other types of POI are considered, each variable representing the number of POI within 3×3 grid cells to include a larger radius, similar to Tran and Ovtracht (2018); Zhao et al. (2019).

Due to benefits for university students, bike-sharing can be observed to be frequently used by students (Tran and Ovtracht 2018). To incorporate this effect, the distance to universities is included. A similar effect can be expected for light rail stations, as there are tariff benefits for long-term ticket holders and in general, a connection between public transit and bike-sharing could be found in previous research (Tran and Ovtracht 2018; Zhao et al. 2019). Closeness to water bodies is assumed to enhance bike-share usage as well (Wang et al. 2015). As a higher population density implies a higher number of potential users, it is also included (Tran and Ovtracht 2018). Other sociodemographic variables are the number of shared flats that might serve as a proxy for the number of students living within grid cells and the age distribution as bike-share usage was found to vary over age (Wang et al. 2015). Here, the age group of people older than 64 years is excluded as it can be deduced from the shares of the other age groups.

The majority of variables is defined as shares (land uses, age groups), ranging from 0 to 1. Distances (in metres) were divided by 1,000 and set to 1 if the closest object of the corresponding category lies farther away than 1,000 m as we assume that higher distances do not make a further difference and this allows us to avoid outliers. Only the numbers of POI, population and shared flats are retained as counts as there is no natural limit. The coordinates are defined in the coordinate system ETRS89/UTM zone 32N.

Methods

In the following, an overview of parametric regression models and their modifications in the modeling of spatial data is given. Then, the general functioning of the XGBoost algorithm is outlined and methods for parameter tuning, model validation, and interpretation are presented.

Parametric regression methods

The relationship between a response variable Y with mean μ and p explanatory variables $X_1, ..., X_p$ is to be modeled. For each observation i, $i = 1, ..., n$, corresponding values of Y and the covariables $\mathbf{x}_i = (x_{i1}, ..., x_{ip}) \in \mathbb{R}^p$ are in the data set.

Out of several different parametric regression models that were fit to the data set, a Tweedie model with spatial splines generated the best results. Therefore, the methods applied in the estimation of this model are presented in detail. The characteristics of the other two approaches used in the regression analyses are displayed in Table 1. For the estimation of Generalized Linear Models (GLMs), the packages `MASS` and `glmmTMB` (in combination with `buildmer` for variable selection) are used.

Tweedie distribution

Let be Y a variable following a Tweedie distribution with Tweedie index parameter $\xi \in \mathbb{R},\ \xi \notin (0, 1)$, scale parameter ϕ, $\mathrm{E}(Y) = \mu$ and variance $\mathrm{Var}(\mu) = \mu^{\xi}$ (Wood 2017, p. 115). If $\xi > 1$, then $\mu > 0$. If $1 < \xi < 2$, the variable Y can also be described as a sum of $N \sim \mathrm{Poi}(\lambda)$ gamma random variables with closed form expressions for λ and the two parameters of the gamma random variables in terms of the Tweedie distribution parameters. If $\xi \leq 1$ or $\xi \geq 2$, let be $\alpha = \frac{2-\xi}{1-\xi}$ and $a(y, \phi, \xi) = \frac{1}{y} \sum_{j=1}^{\infty} W_j$ where

$$\log(W_j) = j \left(\alpha \log(\xi - 1) - \frac{\log(\phi)}{\xi - 1} - \log(2 - \xi) \right) - \log(\Gamma(j + 1)) - \log(\Gamma(-j\alpha)) - j\alpha \log(y).$$

Then, in all other cases, the density of Y can be described as follows (Dunn and Smyth 2005; Wood 2017):

$$f(y) = a(y, \phi, \xi) \exp \left[\frac{\mu^{1-\xi}}{\phi} \left(\frac{y}{1 - \xi} - \frac{\mu}{2 - \xi} \right) \right].$$

Table 1. Characteristics of parametric regression methods

	Poisson	NB
PMF[a]	$f(y) = \frac{\lambda^y}{y!} e^{-\lambda}$	$f(y) = \frac{\Gamma\left(y + \frac{1}{\alpha}\right)}{\Gamma(y+1)\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{1}{1+\alpha\lambda}\right)^{\frac{1}{\alpha}} \left(1 - \frac{1}{1+\alpha\lambda}\right)^{y}$
Parameters	$\lambda > 0$	$\lambda > 0,\ \alpha > 0$
Parameter space	$Y \in \mathbb{N}$	$Y \in \mathbb{N}$
$\mathrm{E}(Y)$	λ	λ
$\mathrm{Var}(Y)$	λ	$\lambda + \frac{\lambda^2}{\theta},\ \theta = \frac{1}{\alpha}$
Link function	log	log
Estimation method	Maximum likelihood	Maximum likelihood
Reference	Hilbe (2011, p. 30f.)	Hilbe (2011, p. 189f.) and Wollschläger (2014, p. 313)

[a] Probability mass function.

Additive models

A GAM is a semi-parametric model that allows to integrate smooth functions of variables as part of the predictor variables (Wood 2017, p. 161). For Y, an exponential family distribution and a smooth monotonic link function g is specified. GAMs are estimated using the package `mgcv`.

Following the notation in Wood (2017, p. 249) we let $\mathbf{A_i}$ denote the ith row of a parametric model matrix, with corresponding parameters $\boldsymbol{\gamma}$, f_j a smooth function of a covariate x_j, and $\mathrm{EF}(\mu_i, \phi)$ an exponential family distribution with mean μ_i and scale parameter ϕ. Then, a GAM can be described by the following model structure:

$$g(\mu_i) = \mathbf{A_i}\boldsymbol{\gamma} + \sum_{j=1}^{p} f_j(x_{ji}),\ y_i \sim \mathrm{EF}(\mu_i, \phi).$$

A smooth function $f \in S$, with S being the space of smooth functions, can be defined as follows. Let be $\{b_1(x), \ldots, b_k(x)\}$ a basis of S and $\beta_1, \ldots, \beta_k$ the unknown parameters (Wood 2017, p. 162). Then

$$f(x) = \sum_{j=1}^{k} b_j(x)\beta_j.$$

Smooth functions allow to examine variables that have a nonlinear effect on the data. A higher k allows higher wiggliness. In this analysis, penalized thin plate regression splines are applied that are based on thin plate splines. To ensure that the model is identifiable, each function sums to zero over the observed variable values.

GAMs are usually fit by penalized iteratively re-weighted least squares (PIRLS) including a penalty for the wiggliness of each smooth function (Wood 2017, p. 249). For smoothing parameter estimation, restricted maximum likelihood (REML) is applied as it allows to estimate all models using the same method in `mgcv`. It is estimated using a Laplace approximation for REML. This method is chosen because of its ability to reduce bias (Wood 2017, p. 83).

For variables that enter a model as a smooth term, multiple coefficients are estimated, one for each basis function. To enable an easier interpretation, the estimated effect of the component smooth functions is plotted on the scale of the corresponding predictor (Wood 2017, p. 183f.). Additionally, confidence intervals as Bayesian credible intervals for the standard errors and the partial residuals are plotted (Wood 2017, p. 293). Partial residuals are defined as the residuals that are produced by the model without the variable under consideration, while keeping all other estimates fixed (Wood 2017, p. 183f.).

Spatial autocorrelation

The nature of the data set as a spatial grid poses the problem of spatial autocorrelation. Due to the first law of geography "everything is related to everything else, but near things are more related than distant things" (Tobler 1970), it can be assumed that there are interdependencies between grid cells in spatial vicinity. This means that the observations are not fully independent. Therefore, models should control for spatial autocorrelation.

In the case of GAM, a two-dimensional smooth term $f(x, y)$ based on the x- and y-coordinates of each grid cell's centroid can be integrated in the model to account for spatial interdependencies (Wood 2017, p. 175). This procedure implies the generalization of x- and y-coordinates as an isotropic coordinate system (Wood 2017, p. 359).

XGBoost algorithm

XGBoost is designed as a scalable machine learning system for gradient boosting that is based on the idea of decision trees. A tree ensemble model is trained applying boosting. Among a range of modeling options, XGBoost can be applied for Poisson regression and Tweedie regression (Chen et al. 2021). It is executed in R using the `xgboost` package.

A decision tree consists of nodes each leading to two different branches, ending in terminal nodes/leaves. Formally, the feature space is subsequently split into disjoint regions $R_1, ..., R_m$, $m \in \mathbb{N}$. For the evaluation of splits, the residual sum of squares (RSS) is used. Starting with the complete data set, in each step, the binary split is chosen that brings the greatest improvement to the RSS of the model (Lesmeister 2017, p. 146). The prediction for a new observation can be calculated as the mean over all y_i in the same terminal node j, $j = 1, ..., m$, with $N_j = |R_j|$ (e.g. Huang et al. 2020):

$$\hat{y}_j = \frac{1}{N_j} \sum_{\mathbf{x}_i \in R_j} y_i.$$

In a tree ensemble model, predictions are estimated based on the explanatory variables using $K \in \mathbb{N}$ additive functions in the following manner (Chen and Guestrin 2016). Let be $\mathcal{F} = \{f(\mathbf{x}) = w_{q(\mathbf{x})}\}(q : \mathbb{R}^p \to T, w \in \mathbb{R}^T)$ the space of regression trees, q the structure of each tree, T the number of leaves in the tree and w the leaf weights, where w_i stands for the score on the ith leaf. Then:

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i),\ f_k \in \mathcal{F}.$$

As a decision tree on its own can suffer from overfitting, measures to lower the variance are taken (Lesmeister 2017, p. 148). In XGBoost, instead of the RSS the following regularized objective is minimized, where l stands for a differentiable convex loss function to measure the difference between observed y_i and predicted values $\hat{y}_i$, $i = 1, ..., n$. $\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2$ is added as a regularization term that penalizes complex models to prevent overfitting, where $\gamma \in [0, \infty)$ and $\lambda \in [0, \infty)$ can be adapted (Chen and Guestrin 2016):

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k).$$

This tree ensemble model is trained in an additive procedure (Chen and Guestrin 2016). Formally, in the tth step, f_t is added to minimize $\mathcal{L}$, where $\hat{y}_i^{(t)}$ represents the prediction determined in the tth iteration.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t).$$

In addition to regularization, shrinkage and column subsampling are applied to avoid overfitting (Chen and Guestrin 2016). In shrinkage, the weights added to the model are scaled by a factor $\eta \in [0, 1]$, the shrinkage parameter, after each step of tree boosting, thus reducing the influence of individual trees. Column subsampling describes the process of building each tree on a subset of all features. Additionally to prevent overfitting, this also enhances computation speed. Next to these critical tuning parameters, further parameters governing the trees' construction can be tuned in the XGBoost implementation (Chen et al. 2021), and we discuss the tuning process in the results section.

Importance plots

Importance plots offer the best visual option to interpret the result of an XGBoost model. It can be distinguished between gain (relative contribution of a model parameter), cover (the relative number of observations that are related to this feature) and frequency (the relative number of how many times a parameter is included in splits of trees). All three measures assume values between 0 and 1, the values for all features sum up to 1. Parameters with higher values are more important for the model (Chen et al. 2021).

Geographically weighted XGBoost

XGBoost does not account for spatial interdependence. Therefore, a GWR is calculated using the geo-referenced predictions made by an XGBoost model as input data, following the approach of Li (2019). The predictions $\hat{y}_i,\ i = 1, ..., n$ are of interest, they are the final predictions of the model. This method is used to include spatial variation and to create smoother predictions. The package `spgwr` is applied for this task.

GWR is implemented using a function K (Kernel function) that describes the spatial relationship of observations and meets the following criteria (Brunsdon, Fotheringham, and Charlton 1998):

- $K(0) = 1$
- $\lim_{d\to\infty} K(d) = 0$
- $K(d_1) > K(d_2), \forall\ d_1 < d_2,\ d_1, d_2 \in \mathbb{R}^+$

K is used to calculate the weight α_{ij} of each observation i to another observation j based on the distance metric d_{ij} between these two observations: $\alpha_{ij} = K(d_{ij})$. In this case, a Gaussian distance-decay based weighting is chosen using the following function, where $h \in (0, \infty)$ stands for the bandwidth (Bivand and Yu 2020):

$$K(d_{ij}) = e^{-(d_{ij}/h)^2/2} = \alpha_{ij}.$$

For each observation $i,\ i = 1, ..., n$, a weight matrix $\mathbf{W}_i$ is calculated whose main diagonal contains the weights that are applied to estimate a weighted regression for this observation. This is used for the estimation of the β-matrix, created column by column from (Brunsdon, Fotheringham, and Charlton 1998):

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}'\mathbf{W}_i\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_i\mathbf{y}.$$

For the selection of an adequate bandwidth h, least squares cross-validation can be applied. The cross-validated sum of squared errors can be defined as a function of h considering the predictions $\hat{y}_i(h)$ determined using this bandwidth: $\text{CVSS}(h) = \sum_{i=1}^{n} (y_i - \hat{y}_i(h))^2$. Here, the ith observation is excluded from the model to prevent that the minimal value lies at $h = 0$. CVSS(h) should be minimized (Brunsdon, Fotheringham, and Charlton 1998).

Error measures

There are several measures that can be applied to compare the quality of two different models. The Akaike Information Criterion (AIC) is used for feature selection in parametric regression models, while the sMAPE and RMSE are calculated to compare the predictions of the parametric regression and the XGBoost model. As the literature does not agree on a single best performance metric, two regularly applied metrics are selected for this task (Botchkarev 2018).

Symmetric mean absolute percentage error

The sMAPE can be calculated as follows for a data set including n observations y_i and predictions $\hat{y}_i,\ i = 1, ..., n$. c is a small positive constant (Li and Axhausen 2019), and we use $c = 1$ here:

$$\text{sMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i + \hat{y}_i + c}.$$

This definition slightly differs from the original definition of this measure by Armstrong (1985). First, the denominator is not divided by 2 to allow for an easier interpretation. Second, following Li and Axhausen (2019), the constant c is included to prevent division by zero that could otherwise occur if both observed and predicted value equal zero. The sMAPE is a dimensionless, percentage-error metric (Botchkarev 2018). Smaller values indicate a better fit.

Results

In the following, the results of the analysis are presented, starting with a descriptive analysis of the data set. Then, the chosen parametric regression model and the XGBoost model are characterized, including residuals and predictions. Finally, both models are compared.

Descriptive analysis

The mean number of trips per grid cell is 8.52. For 3,218 of 8,955 grid cells, 0 trips were observed, the maximum number is 329. There are several upper extreme values and the distribution of observed values is right-skewed.

The spatial distribution of the observed trip counts for each grid cell is depicted in Figure 1. It becomes apparent that many grid cells with zero trips appear clustered in the outer areas of the study area, especially in the northern part. In the central areas, trip numbers are rather high on average. Apart from that, several large clusters of grid cells with higher values are visible that are dispersed throughout the study area or lie along certain lines. Single cells where higher numbers of trips could be observed appear infrequently also in the outer areas.

The spatial distribution of all measured values per grid cell in the training data is presented in Figure 2 and the one of the test data in Figure 3. These figures indicate that the spatial distribution of trip counts in the total data could be preserved in both data sets. The mean number of trips per grid cell is 4.31 in the training data set and 4.21 in the test data set, the maximum counts lie at 167 and 168, the number of grid cells where 0 trips started are 3,878 and 3,954, respectively.

Independent variables

The main metrics of the independent variables used in the modeling of the trip numbers per grid cell are presented in the following Table 2. All variables assume nonnegative values, most of them are continuous. The variables representing the number of different POI nearby are count variables and therefore discrete. Still, most POI occur rather infrequently causing all mean values to lie below 1. Only *shops, food outlets, bars* amounts to a much higher mean of 5.55. The variables *population* and *shared flats* are discrete as well. *x*- and *y-coordinate* are measured as UTM coordinates and can be treated as continuous in the analysis.

Independent variables that exhibit the strongest correlations with the *trip count* regarding the Bravais-Pearson correlation coefficient are *shops, food outlets, bars* (0.40), *university* (−0.32), *shared flats* (0.31), *buildings* (0.30), *population* (0.28), *light rail* (−0.26), *FNP 1* (0.23), *hotels* (0.22), and *18–29 years* (0.21).

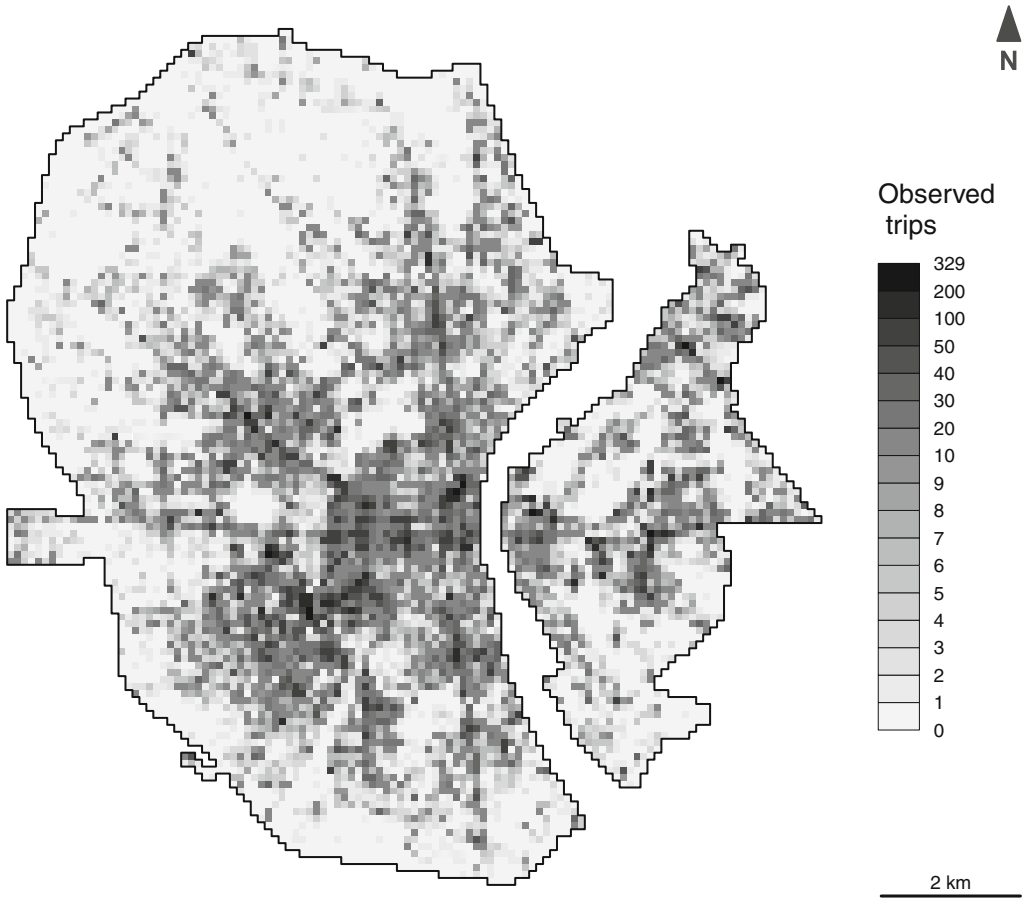

Figure 1. Spatial distribution of the number of trips per grid cell.

High correlations between independent variables can be avoided to a large part. The strongest correlations occur especially between *shops, food outlets, bars* and a small number of other variables. The two highest correlations lie at 0.61 and 0.57, the correlations of 12 other variable combinations lie between 0.4 and 0.5. But all in all, correlations between the independent variables are rather low, most of them lie between −0.2 and 0.2, a majority even between −0.1 and 0.1.

Parametric regression model

In the following, the different metrics resulting from the various parametric regression models that were considered in the process are presented. Based on this, the model producing the most adequate predictions of the test data set is chosen and presented in further detail.

For the selection of parameters to be included in the models, two different approaches are used. As 41 possible parameters (plus in some cases x- and y-coordinate) were assessed in the different models, parameters are selected through stepwise forward selection minimizing the AIC for the models estimated in `MASS` and `glmmTMB`.

When estimating a model with smooth parameters in `mgcv`, it is possible to add an extra penalty that penalizes functions in the smoothing penalty null space (Wood 2017, p. 214).

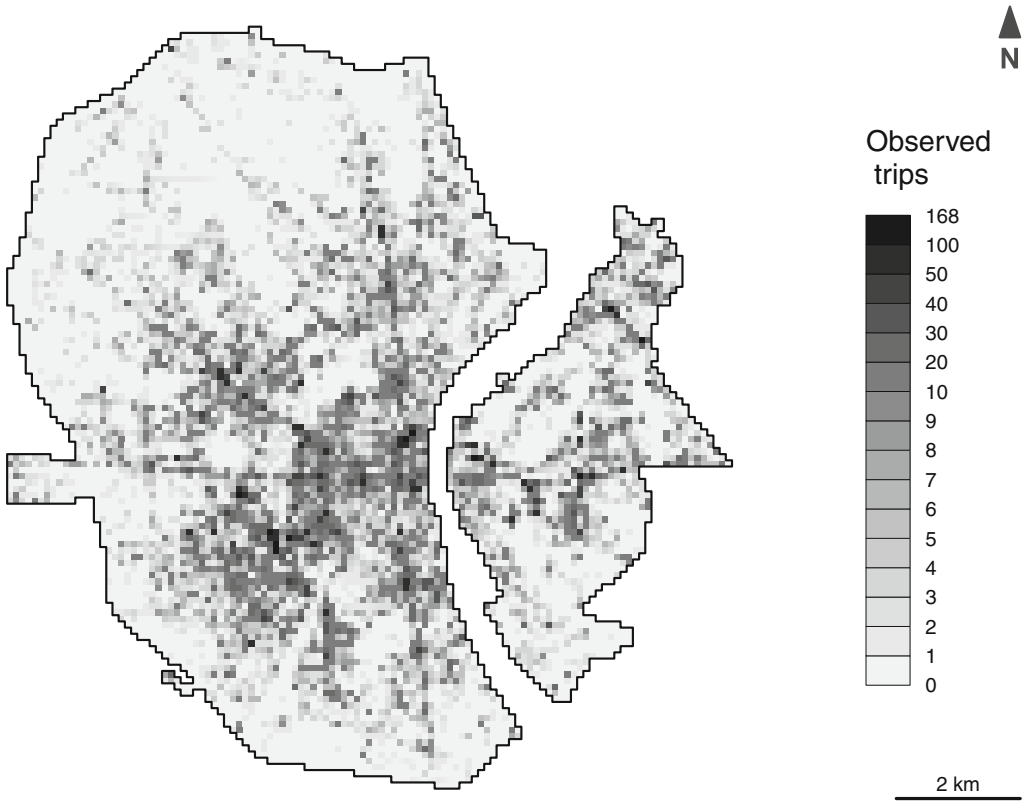

Figure 2. Spatial distribution of the number of trips per grid cell in the training data set.

This allows to completely exclude single parameters. Therefore, the models estimated using the package `mgcv` use smooth terms for all variables to enable variable selection. To choose an adequate value for the number of basis dimensions k for each variable, multiple models including all possible variables are estimated starting with $k = 3$ for all. Then, k is successively increased until sufficient. In the final models, $k = 400$ is defined as the upper limit for the interaction of both coordinates, $k = 15$ for the variables *shops, food outlets, bars*, and *light rail*, $k = 10$ for most other variables and $k = 3$ for variables that assume 20 or less unique values. Then, the influence of all variables on the model is evaluated as a combination of significance in the model summary, effective degrees of freedom (edf) and the structure of the partial residual plot. Those variables that are insignificant, have low edf values < 1 and/or where the partial influence plot shows a straight line at 0, are removed from the full model. This allows to improve the AIC even more. The resulting AIC, RMSE and sMAPE of all six parametric regression models considered are presented in the following Table 3.

The comparison of the three different measures does not lead to a clear decision: The NB models with zero inflation or spatial splines and the Tweedie model with spatial splines result in the lowest AIC values, while the Poisson model with zero-inflation and spatial splines leads to a far higher AIC in comparison to all other models. Nevertheless, the latter model is rather successful at predicting the values of the test data set regarding the RMSE, which is the lowest of all models. The second best RMSE is generated by the Tweedie model with spatial splines, while the NB model with spatial splines produces the third best. The RMSE of the NB model in

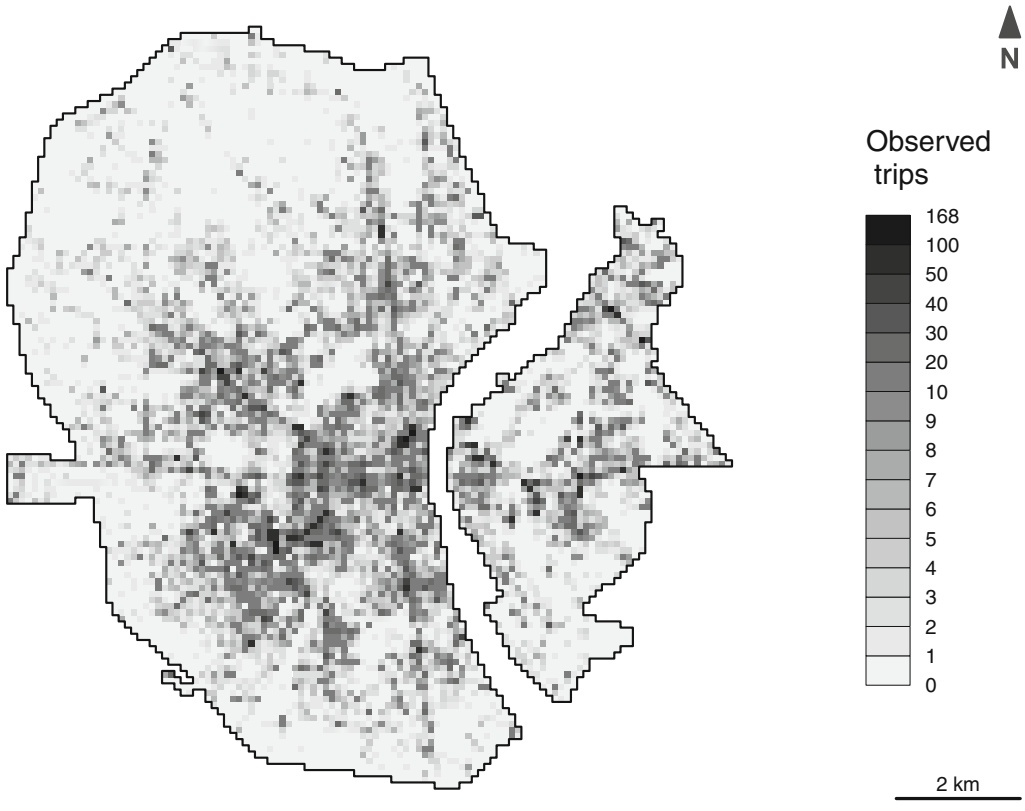

Figure 3. Spatial distribution of the number of trips per grid cell in the test data set.

MASS that performs well considering AIC lies much higher than those of all other models. Also regarding sMAPE, the three models estimated in mgcv perform best, all leading to sMAPE = 0.33. The three other models generate similar values as well, but considerably higher (sMAPE = 0.39/0.40). The different ratings indicate the importance of using more than one performance measure.

All three measures combined, the three models estimated in mgcv outperform the models estimated in MASS and glmmTMB. Both the zero-inflated Poisson model and the NB model with spatial splines lead to an especially high value for one measure (AIC and RMSE, respectively), while the Tweedie model performs consistently well considering all three measures. Therefore, this model is chosen to be compared to the XGBoost model and will be presented in further detail in the following section.

Model results

The parameters estimated for the selected variables in the Tweedie model with spatial splines in mgcv are specified in Table 4. As this research focuses on the comparison between this model and the XGBoost model, the interpretation of model parameters focuses mainly on the plausibility of the selected parameters' influences on bike-sharing usage.

For the intercept, the usual regression output is displayed, while the output differs for smooth terms. For these variables, edf and test statistics used in an Analysis of Variance (ANOVA) test to test the significance of the smooth are displayed. The edf indicate the complexity of a smooth.

Table 2. Minimum, Mean $\bar{x}$, Maximum and SD s for the Independent Variables Used in the Analysis, for All $n = 8{,}955$ Grid Cells

Variable	Min.	Mean	Max.	SD
Trip count	0.00	8.52	329.00	18.01
Train trip count	0.00	4.31	167.00	9.19
Test trip count	0.00	4.21	168.00	9.06
FNP 0	0.00	0.29	1.00	0.40
FNP 1	0.00	0.09	1.00	0.26
FNP 2	0.00	0.05	1.00	0.18
FNP 3	0.00	0.09	1.00	0.25
FNP 4	0.00	0.02	1.00	0.14
FNP 8	0.00	0.01	1.00	0.07
FNP 9	0.00	0.00	1.00	0.03
FNP 11	0.00	0.03	1.00	0.12
FNP 12	0.00	0.05	1.00	0.18
FNP 13	0.00	0.05	0.99	0.12
FNP 15	0.00	0.01	1.00	0.06
FNP 16	0.00	0.01	1.00	0.08
FNP 17	0.00	0.04	1.00	0.17
FNP 18	0.00	0.00	0.60	0.01
FNP 21	0.00	0.00	0.97	0.02
Green spaces	0.00	0.16	1.00	0.30
Buildings	0.00	0.20	1.00	0.18
Shops, food outlets, bars	0.00	5.55	204.00	15.14
Health care facilities	0.00	0.17	14.00	0.67
Schools	0.00	0.17	5.00	0.49
Kindergartens	0.00	0.38	5.00	0.68
Museums	0.00	0.04	5.00	0.27
Event venues	0.00	0.08	5.00	0.36
Libraries	0.00	0.03	4.00	0.23
Public institutions	0.00	0.02	4.00	0.16
Sports facilities	0.00	0.11	7.00	0.40
Tourist attractions	0.00	0.21	10.00	0.77
Hotels	0.00	0.06	4.00	0.30
Places of worship	0.00	0.04	3.00	0.22
Playgrounds	0.00	0.15	4.00	0.44
University	0.00	0.78	1.00	0.32
Bus	0.00	0.24	1.00	0.20
Light rail	0.00	0.38	1.00	0.28
Arterial road	0.00	0.14	1.00	0.19
Water	0.00	0.68	1.00	0.33
Population	0.00	61.22	581.00	84.29
Shared flats	0.00	14.38	147.00	18.70
0–17 years	0.00	0.10	1.00	0.09
18–29 years	0.00	0.13	1.00	0.12
30–49 years	0.00	0.26	1.00	0.19
50–64 years	0.00	0.12	1.00	0.10
x-coordinate	350,164.68	355,322.69	361,764.68	2,659.74
y-coordinate	5,639,809.24	5,645,988.98	5,652,009.24	2,992.29

Table 3. AIC, RMSE und sMAPE of the Parametric Regression Models

Model (package)	AIC	RMSE	sMAPE
NB model (`MASS`)	37454.08	9.17	0.39
Tweedie model (`glmmTMB`)	38473.29	7.82	0.40
NB model with zero-inflation (`glmmTMB`)	37086.04	8.36	0.39
Poisson model with zero-inflation and spatial splines (`mgcv`)	46736.67	6.05	0.33
NB model with spatial splines (`mgcv`)	35768.09	7.63	0.33
Tweedie model with spatial splines (`mgcv`)	36401.62	6.58	0.33

Table 4. Estimate, SE, *t* value and *P*-Value for the Intercept and edf, ref.df, *F* and *P*-value for All Variables Included in the Model as Smooth Terms

Variable	Estimate	SE	*t* value	*P*-value
Intercept	0.53	0.02	27.85	<0.01
Variable	edf	ref.df.	*F*	*P*-value
FNP 0	3.21	9.00	4.56	<0.01
FNP 2	5.43	9.00	5.05	<0.01
FNP 3	1.45	9.00	1.36	<0.01
FNP 8	2.00	9.00	0.98	<0.01
FNP 12	6.56	9.00	8.03	<0.01
FNP 13	4.13	9.00	4.27	<0.01
FNP 15	1.76	9.00	3.39	<0.01
FNP 17	3.92	9.00	6.48	<0.01
Buildings	6.52	9.00	10.32	<0.01
Tourist attractions	0.95	2.00	9.10	<0.01
Shops, food outlets, bars	8.54	14.00	10.45	<0.01
University	7.48	9.00	8.46	<0.01
Arterial road	5.85	9.00	2.99	<0.01
Bus	7.07	9.00	21.94	<0.01
Light rail	12.01	14.00	22.24	<0.01
Water	5.63	9.00	3.40	<0.01
Green spaces	5.89	9.00	9.45	<0.01
Population	2.40	9.00	14.59	<0.01
18–29 years	2.91	9.00	4.06	<0.01
x- and *y*-coordinate	215.10	399.00	3.65	<0.01

If edf = 1, the smooth can be represented as a straight line, modeling a linear term, while higher values indicate a higher wiggliness. The reference degrees of freedom (ref.df) used in computing test statistics represents $k - 1$, k being the number of basis dimensions specified in the model. *F* and the corresponding *P*-value represent the results of the ANOVA test.

In the variable selection process, the following were excluded from the model: *FNP 1, FNP 4, FNP 9, FNP 11, FNP 16, FNP 18, FNP 21, health care facilities, schools, kindergartens, museums, event venues, libraries, public institutions, sports facilities, hotels, places of worship, playgrounds, shared flats, 0–17 years, 30–49 years*, and *50–64 years*. All smooth terms are

individually significant at a 95% confidence level, which underlines the value of the selected variables. The edf vary strongly between variables, ranging from 0.95 (*tourist attractions*) to 215.10 (*coordinates*). The high edf estimated for the coordinates indicate the presence of an influence factor that has been left out of the model so far or alternatively that there are strong spatial correlations between the observations. Other variables with high values for edf are *FNP 2* (5.43), *FNP 12* (6.56), *buildings* (6.52), *shops, food outlets, bars* (8.54), *university* (7.48), *arterial road* (5.85), *bus* (7.07), *light rail* (12.01), *water* (5.63) and *green spaces* (5.89). In combination with the significant *P*-values, these indicate a nonlinear influence of the variables on bike-sharing.

The direction of these relationships can be further interpreted visually through plots that display the partial effects for a smoothed variable. 95% pointwise confidence intervals and partial residuals are included as well. The partial effect plots of all variables are included in Appendix S1.

To assess whether k is chosen adequately for each variable, the k are compared to the edf as the edf values should be considerably lower. Here, Table 4 can be consulted again. None of the estimates for edf comes close to the value of the corresponding k suggesting that all k are sufficiently high. A visual interpretation of the partial effect plots underlines the adequacy of the selected k, as no partial residuals indicate that a higher wiggliness ist needed.

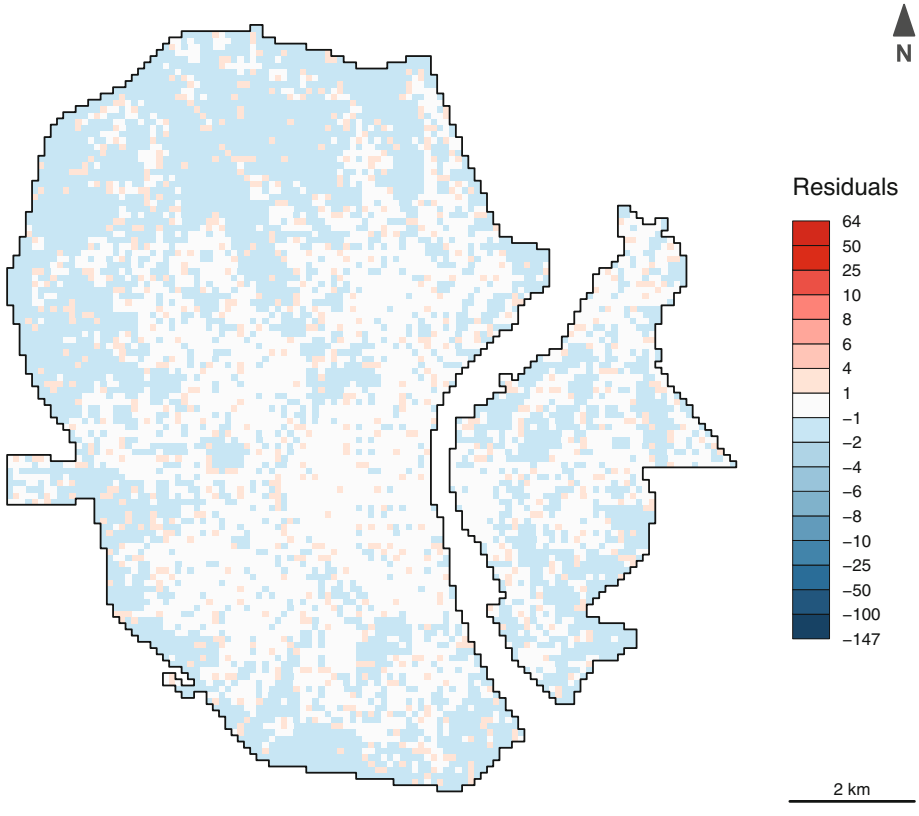

Figure 4. Residuals in the Tweedie GAM with spatial splines.

In Figure 4, the residuals of the fitted values for the number of bike-sharing trips per grid cell in the training data set are displayed spatially, each grid cell representing the residual value calculated for this observation. As all residuals lie between −1.49 and 2.82, the observations of the training data set can be modeled quite well. In total, the conditional distribution of the model is still overdispersed, even though variables for points of interest are included. But in comparison to an intercept only model, a high share of the variance and overdispersion can be captured by the covariates. Especially in the central parts, high numbers are successfully estimated. The residuals appear in a mixed structure and the absolute value of most residuals is less than 1. In the outer areas, where 0 trips occur in most grid cells, the number of trips is often overestimated.

Predictions

The predicted number of bike-sharing trips per grid cell based on the data set is shown in Figure 5. Here, it becomes apparent that visually, there are strong similarities between the predictions and the observed values depicted in Figure 3. Especially in the central areas, high numbers of trips (maximum: 92) are predicted, while zero trips are predicted in the outer areas, especially in the northern part.

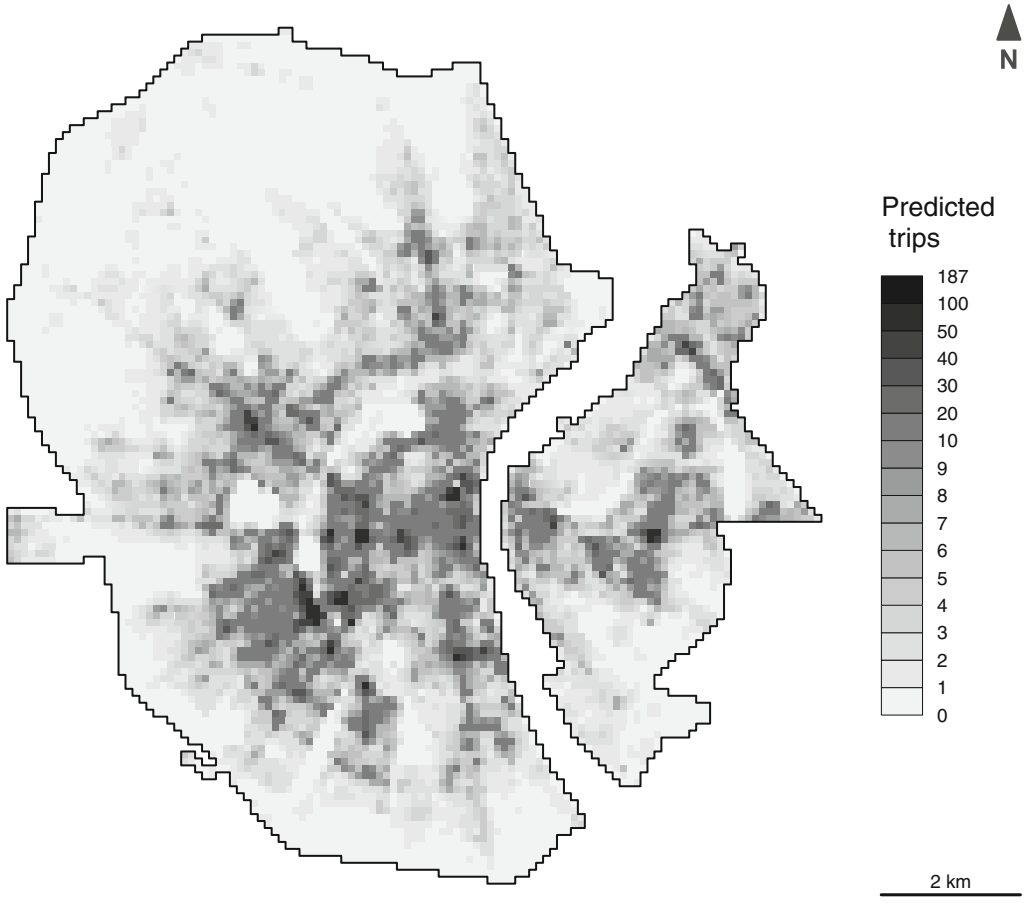

Figure 5. Predictions of bike-sharing trip counts determined by the Tweedie GAM with spatial splines.

All in all, the chosen GAM, a Tweedie regression with spatial splines estimated in `mgcv`, performs well at modeling the test data set and offers informative means of interpretation.

XGBoost model

One of the major advantages of XGBoost, its flexibility to be applied on a wide range of data sets, leads to the necessity of hyperparameter tuning to adjust the model to the specific data set. It is possible to estimate XGBoost models without hyperparameter tuning using the default values instead, but the adjustment leads to much more appropriate models (Ryu, Shin, and Chung 2020). While there are over 30 hyperparameters in XGBoost, it is sufficient to tune a much smaller number, outlined in the following. Further descriptions of the parameters that are tuned are included in Appendix S1.

As general model settings, `booster = "gbtree"`, `eta` = 0.3, `eval_metric = "rmse"` in combination with `maximize = FALSE` and `early_stopping_rounds` = 10 are chosen.

Both objectives `count:poisson` and `reg:tweedie` are included in the analysis to determine whether a Poisson or Tweedie model generates more accurate results. For each objective the same two-step process is applied: first, the optimal number for `nrounds` is determined through cross-validation using the default values for the XGBoost model. The data is randomly split into five equal-sized subsamples and the RMSE is aimed to be lowered as much as possible. This leads to `nrounds` = 57 for the Poisson model and `nrounds` = 36 for the Tweedie model. Using these values in combination with the default parameters, two models are fit leading to an RMSE of 3.63 (training data), 3.99 (test data) for the Poisson model and 4.28 (training data), 4.46 (test data) for the Tweedie model.

The parameters `max_depth`, `gamma`, `subsample`, `colsample_bytree`, and `min_child_weight` are optimized in a second step applying a grid search approach. Besides random search this method is frequently used for hyperparameter optimization (Putatunda and Rama 2018). While random search estimates a large number of models using randomly chosen parameters, in grid search a grid of all combinations of defined values for the parameters is created and for each combination a model is estimated. This allows to select the hyperparameters of the best model. For the parameters to be optimized, the following values are considered, mostly based on Huang et al. (2020):

- `max_depth` $\in \{3, 8, 13, 18, 23, 28, 33, 38\}$
- `gamma` $\in \{0, 0.05, 0.1, 0.15, 0.2\}$
- `subsample` $\in \{0.5, 0.6, 0.7, 0.8, 0.9\}$
- `colsample_bytree` $\in \{0.5, 0.6, 0.7, 0.8, 0.9\}$
- `min_child_weight` $\in \{0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$

This combination of parameter values requires the estimation of 11,000 models. Additional parameters could be tuned as well, but as the tuning of the chosen set of hyperparameters already leads to sufficiently small RMSE values, tuning is limited to these. As the RMSE values produced in the grid search for the Poisson models are much smaller on average than those of the Tweedie models, the Poisson model is chosen. This process leads to the following parameters for the Poisson model minimizing the training data RMSE that reach the values RMSE = 2.97 (and an extremely low RMSE = 0.10 for the training data) and sMAPE = 0.19 for the test data: `max_depth` = 33, `gamma` = 0, `subsample` = 0.9, `colsample_bytree` = 0.5, `min_child_weight` = 0, `nrounds` = 57. Additionally, in `xgb.train()`, the parameter

`early_stopping_rounds` is set to 10, causing the model training to stop if the RMSE has not improved in the last 10 rounds. This measure will potentially speed up the fitting process.

One reason for the extremely low RMSE that can be reached by the model may lie in the spatial structure of the data, where each grid cell can be identified by its unique combination of *x*- and *y*-coordinate. A machine learning model could theoretically achieve a perfect fit if it links these combinations to the values of the dependent variable. Several measures are taken to prevent this. First, XGBoost includes different types of subsetting (`subsample` = 0.9, `colsample_bytree` = 0.5), and single trees are limited in depth (`max_depth` = 33).

Model results

In the following, the importance plots of the features included in the model are illustrated to allow an interpretation regarding the impact of each variable. First Figure 6 represents the gain scores of each variable. Here, it is visible that the variables *shops, food outlets, bars*, and *light rail* contribute most to the model accuracy, followed by *university* and *buildings*. Regarding the

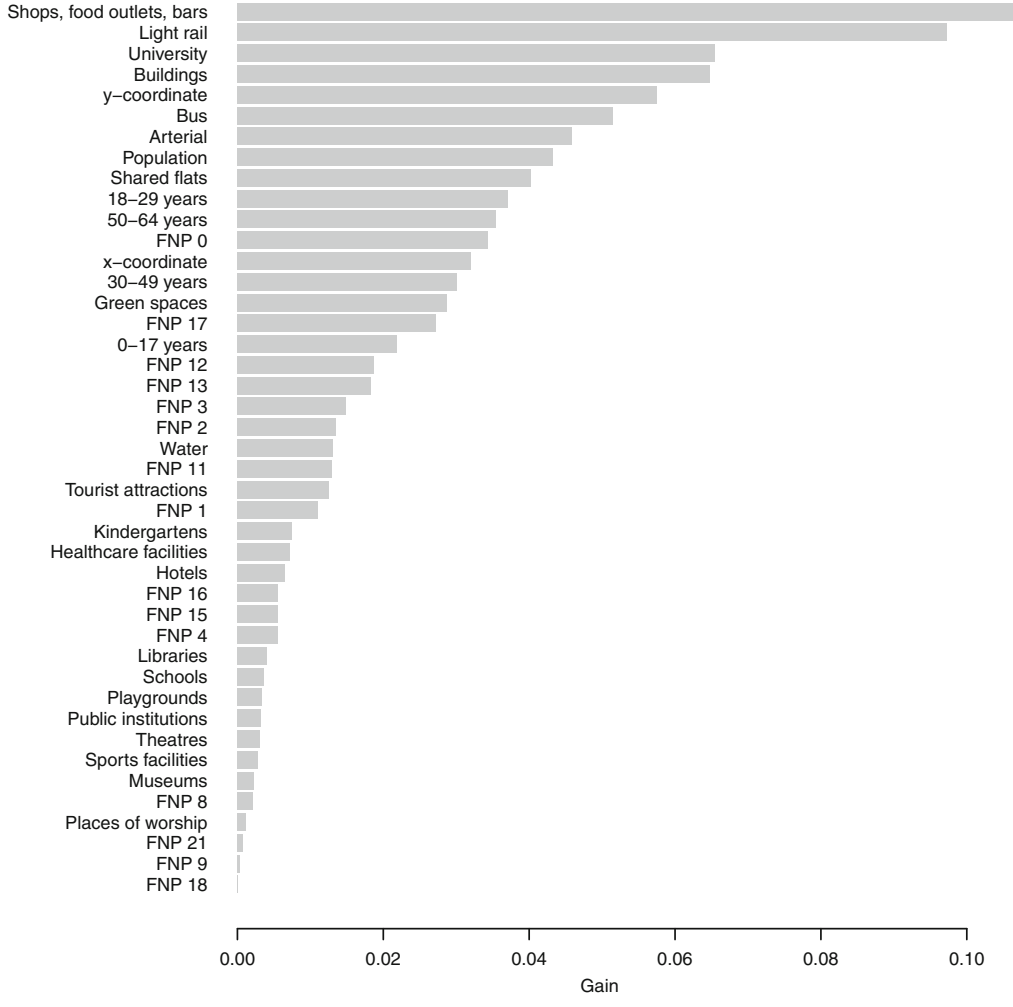

Figure 6. Importance plot of the features in the XGBoost model – gain.

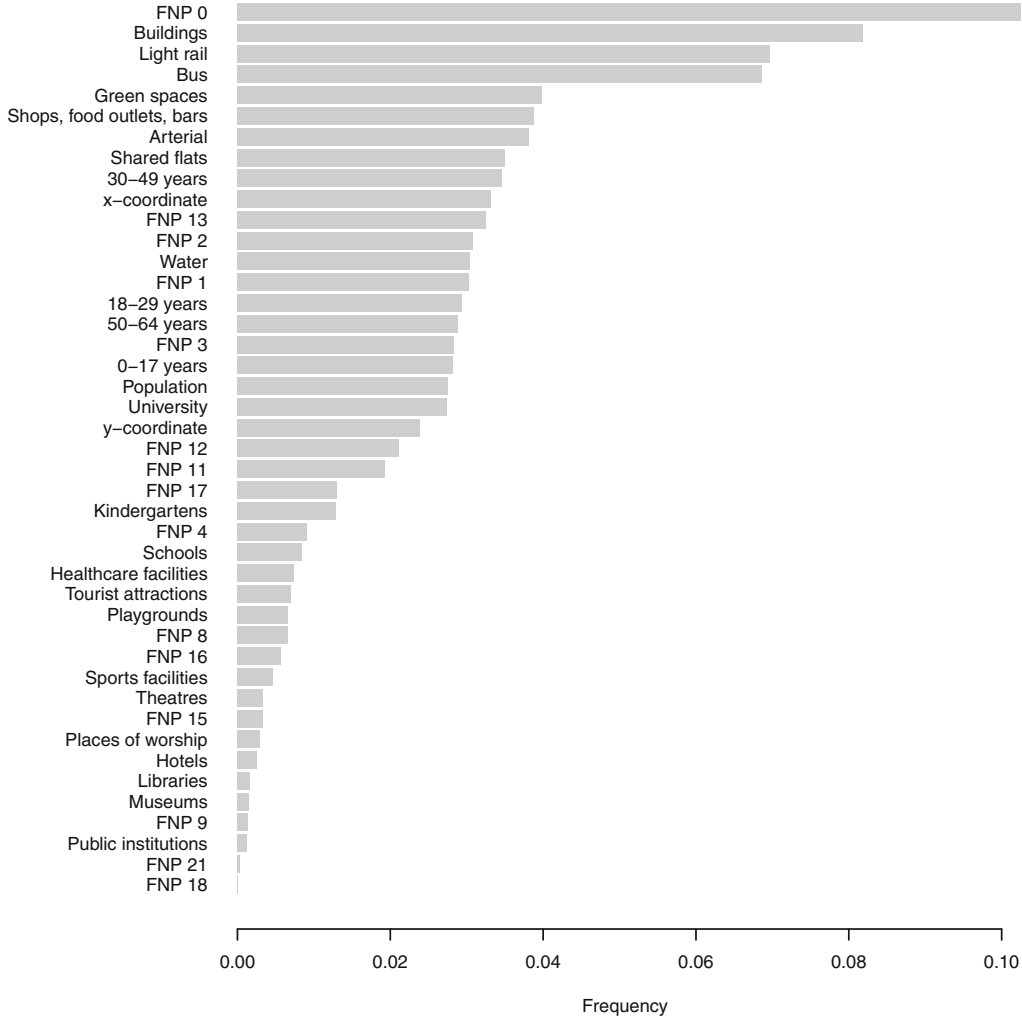

Figure 7. Importance plot of the features in the XGBoost model – frequency.

frequency of each variable to be used in trees shown in Figure 7, the plot exhibits a slightly different structure: the variable *FNP 0* representing the share of residential building land is used most often, followed by *buildings, light rail* and *bus*. All other variables contribute considerably less to the model. Figure 8 presents the number of observations that are related to the variable considered. According to this measure, *light rail* is the variable with the highest score, followed by *x-coordinate, 30–49 years, 50–64 years, bus* and *buildings*. Several other variables exhibit only slightly lower scores.

Although the most important variable is a different one for each plot, it becomes apparent that some variables reach high scores in all three methods to assess importance. All three plots combined, especially the variables *light rail, buildings* and *bus* rank among the highest ones, *arterial road* and *green spaces* rank consistently high as well. Some other variables, such as *shops, food outlets, bars, university, y-coordinate, population* reach extremely high values in one category and rank in the middle for the other two. Especially regarding the variable *FNP 0* it

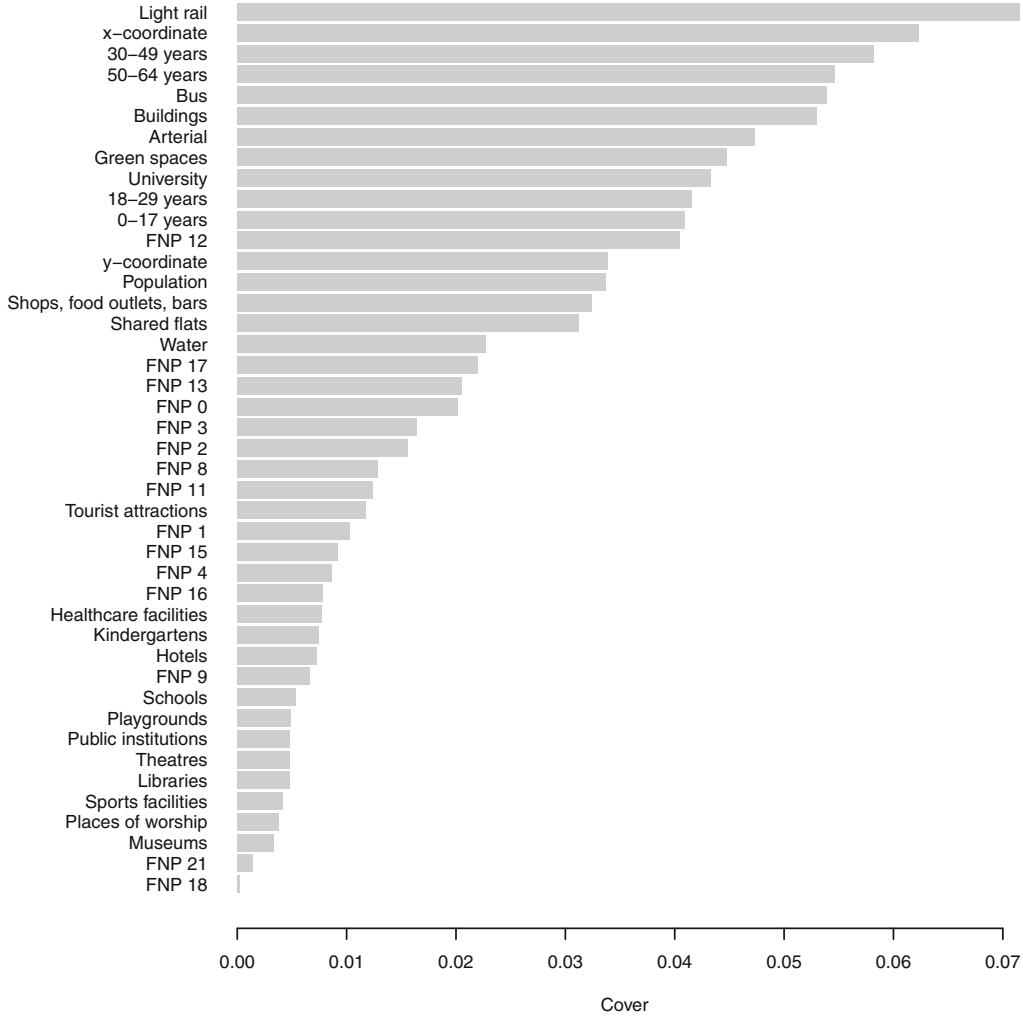

Figure 8. Importance plot of the features in the XGBoost model – cover.

becomes apparent that the evaluation of importance is not consistent through all three plots: While it ranks as the variable with the highest importance by far regarding frequency (Figure 7), it reaches only much lower ranks considering the other two measures.

In all three plots, the variable that is of smallest importance for the model is *FNP 18*, the share of a certain type of special building land, contributing close to nothing to the model. Other variables that rank very low are *FNP 21, FNP 9, museums, places of worship, sports facilities* and *public institutions*. This finding matches to the result of the Tweedie GAM, where *FNP 18* is the first variable to be excluded from the model and all other variables mentioned are removed as well. Most other variables that are excluded in the GAM reach values in the bottom and middle ranges in the importance plots. But also differences become apparent that illustrate the differences between both methods: The variables *shared flats* and *0–17 years, 30–49 years, 50–64 years* that are excluded in the GAM reach high importance values in XGBoost.

Similar to the results of the parametric model and in line with the previous research (Tran and Ovtracht 2018; Zhao et al. 2019), especially those variables that describe the transport related infrastructure of Cologne seem to have a strong influence on the model construction in XGBoost (*light rail, bus, arterial road*). The most important other influence factors are the distance to universities (*university*), the number of certain POI (*shops, food outlets, bars*), the population density (*population*), the share of built-up space (*buildings*) and green spaces (*green spaces*) per grid cell. These relationships could also be found in other studies, such as Tran and Ovtracht (2018). The coordinate values (*x-coordinate, y-coordinate*) highly add to the model as well, especially regarding gain and cover. Still, in all three importance plots there are variables ranking higher than the coordinates, meaning the model does not simply rely on the spatial positions to model the data but actually takes into account a wide range of different features.

Predictions

The structure of the predictions of bike-sharing trips per grid cell determined by the XGBoost model appears to be similar to those generated by the Tweedie GAM and the observations themselves (figure included in Appendix S1). Still, the predictions reach much higher values than in the GAM (maximum: 166). In the outer parts of the study area, zero trips are predicted for most grid cells. The variance of predictions for grid cells close to each other in the central parts seems to be rather high as frequently grid cells with very high predictions lie next to those with very low ones. This is the most noticeable difference to the predictions of the GAM, where usually similar values are generated for grid cells neighboring each other.

Geographically weighted XGBoost

Using the predictions generated in XGBoost as input values, a GWR is performed. Here, an optimal bandwidth of 51.45 (m) can be determined. In comparison to the RMSE = 2.97 and sMAPE = 0.19 that can be achieved with XGBoost, the predictions generated by geographically weighted XGBoost lead to RMSE = 3.72 and sMAPE = 0.24. This indicates that the general accuracy of predictions cannot be further increased through the estimation of GWR.

The final predictions of the geographically weighted XGBoost model are displayed in Figure 9. The structure of these predictions is naturally very similar to those of the previous XGBoost, as they are directly based on them. It is visible that the GWR could create smoother predictions, that, visually, appear even more comparable to the predictions generated by the Tweedie regression. GWR is successful at lowering the variance of predictions of neighboring grid cells.

In conclusion, geographically weighted XGBoost decreases the predictive accuracy of the already extremely well fit XGBoost model. Still, it succeeds at creating more realistic predictions and lowers variance of predictions in close vicinity.

Comparison

In the following, the outcomes and the estimation process of both modeling approaches are compared regarding predictive accuracy, time efficiency and ease of interpretation.

In Figure 10, we compare all models estimated in this research. Here, the high accuracy of XGBoost as well as the advantage of models including spatial splines in contrast to other parametric regression models become apparent. For both models examined in detail, RMSE and sMAPE combined with time needed for computation, in the case of XGBoost split into parameter tuning and computation of the model itself, are contrasted in Table 5. While the value

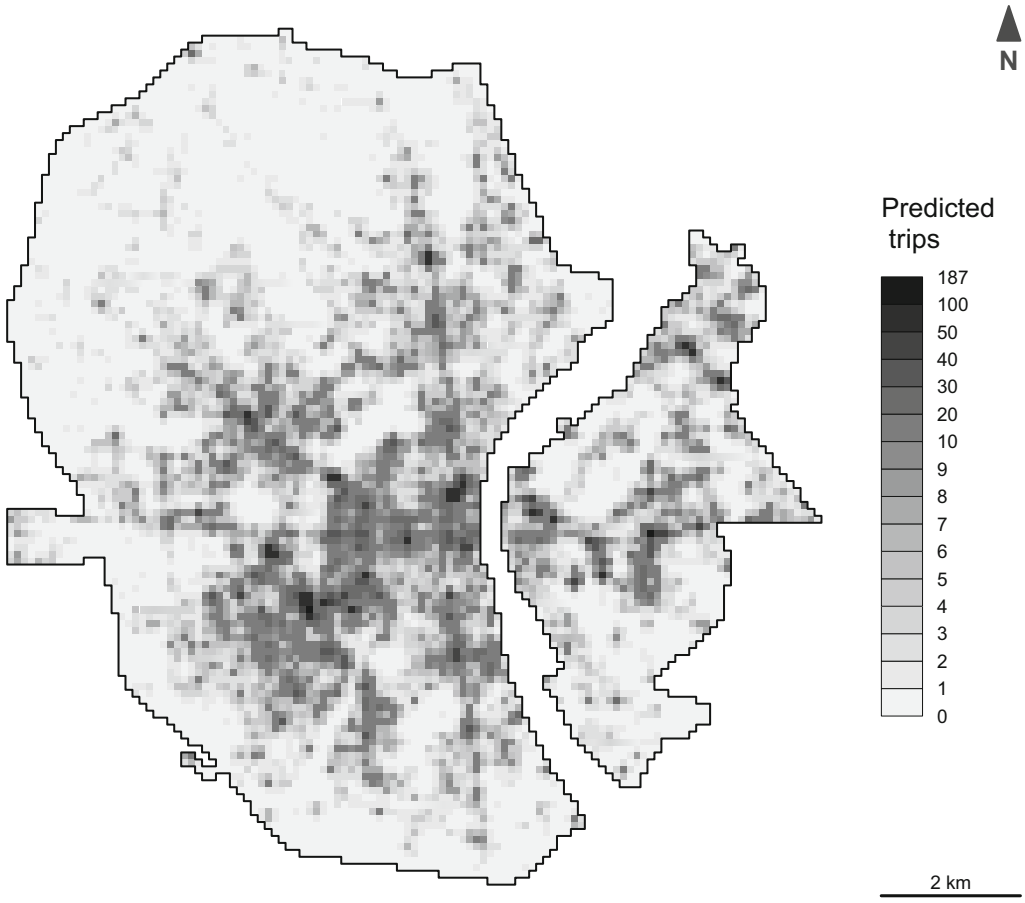

Figure 9. Predictions of bike-sharing trip counts determined by geographically weighted XGBoost.

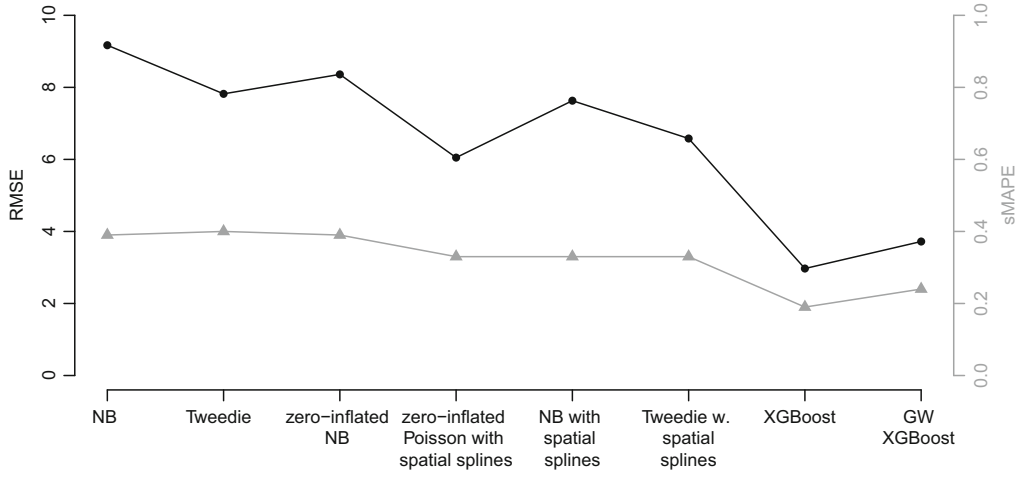

Figure 10. RMSE and sMAPE of all models considered.

Table 5. RMSE, sMAPE and Computation Time for Both Models

	Tweedie GAM	XGBoost	Geographically weighted XGBoost
RMSE	6.58	2.97	3.72
sMAPE	0.33	0.19	0.24
Computation time	1:50:46.44 h	0:00:02.40 h	0:03:55.77 h
Computation time – hyperparameter tuning	-	3:23:32.56 h	-

for the Tweedie GAM lies at 6.58 and the lowest value among all parametric regression models at 6.05, even an XGBoost model estimated using the default parameter settings reaches a lower RMSE of 3.99. After hyperparameter tuning, it is even possible to lower the RMSE to 2.97. Considering computation time, the estimation of a single XGBoost model takes far less time (model with selected parameters: 2.40 s, with default parameters: 0.33 s) than the estimation of a GAM with the selected data set and settings (1:50:46.44 h). In contrast, the hyperparameter tuning includes the estimation of many more models, in this case 11,000 and therefore requires a much higher time (3:23:32.56 h). The estimation of a GWR based on the model predictions of XGBoost only requires an additional time of 3:55.77 min including the search for an optimal bandwidth and estimation of the model. Keeping the lower RMSE of XGBoost in the model with default parameters in mind, (geographically weighted) XGBoost is the superior model in terms of computation time only when parameter tuning is omitted, say by defaulting to parameters from an older version of the same data set, or a less elaborate tuning rationale is used than here.

Predictions

In Figure 11, the differences between the predicted values by the Tweedie GAM and the observed values in the test data set are shown. The greatest overestimation of trips by the model is 63.61, the greatest underestimation 146.59. Especially the high accuracy in the outer areas is visible, where the predictions of very small numbers of trips are mostly correct. In general, the spatial distribution of overestimations shows large similarities to the distribution of predictions themselves. This means that especially in the central parts of the study area, there is a visible clustering of grid cells where higher numbers of trips are predicted but the number of trips is overestimated by the model. In contrast, grid cells where the number of trips is underestimated are spread out more evenly throughout the research area.

In comparison to the results of the Tweedie GAM, a different structure becomes apparent regarding the differences between the predictions of XGBoost and the observed numbers of trips in Figure 12. Differences are much smaller ranging only from −41.20 to 38.59 and the areas where predictions are correct are larger. Still, the largest areas where the predictions are correct lie in the outer areas, especially in the northern part of the research area. Here, 0 trips were observed in most grid cells.

The differences between the predictions of the geographically weighted XGBoost model and the observations visible in Figure 13 illustrate why both measures RMSE and sMAPE assume higher values for this method in comparison to XGBoost. Differences generally appear to be higher and the areas where estimations are correct seem to be slightly smaller. The minimum

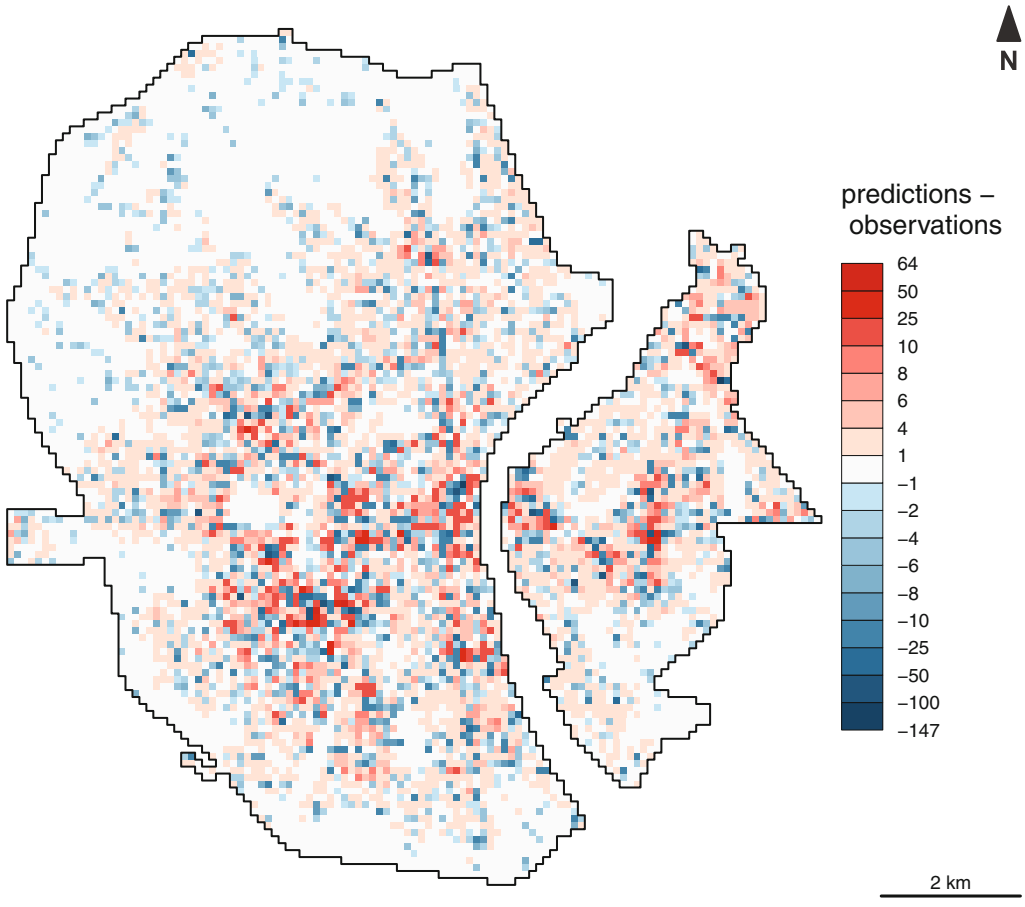

Figure 11. Difference between predicted and observed values of the test data set, predictions determined by the Tweedie GAM with spatial splines.

value lies far lower than the one of XGBoost, at −79.99, while the maximum value is even much smaller and only reaches 17.89.

Spatial autocorrelation

As this analysis deals with a spatial data set, the abilities of each model to handle the specific requirements connected to the issue of spatial autocorrelation can be evaluated. In GAM, the interaction between the *x*- and *y*-coordinates can be included through a spline modeling an interaction term. This also allows to interpret the relation of the dependent variable to the locations in space. XGBoost allows to include x- and y-coordinates as variables but does not include a method to explicitly deal with spatial locations. Therefore, an additional GWR is performed in this analysis that increases the RMSE and sMAPE reached by the model, lowering the predictive accuracy but creating smoother predictions. In conclusion, GAM offers a method to include spatial information directly in the estimation of the model while XGBoost fails to do so. Additionally, in XGBoost the inclusion of coordinates in the model involves the risk of XGBoost fitting the model only to the coordinates if the settings allow.

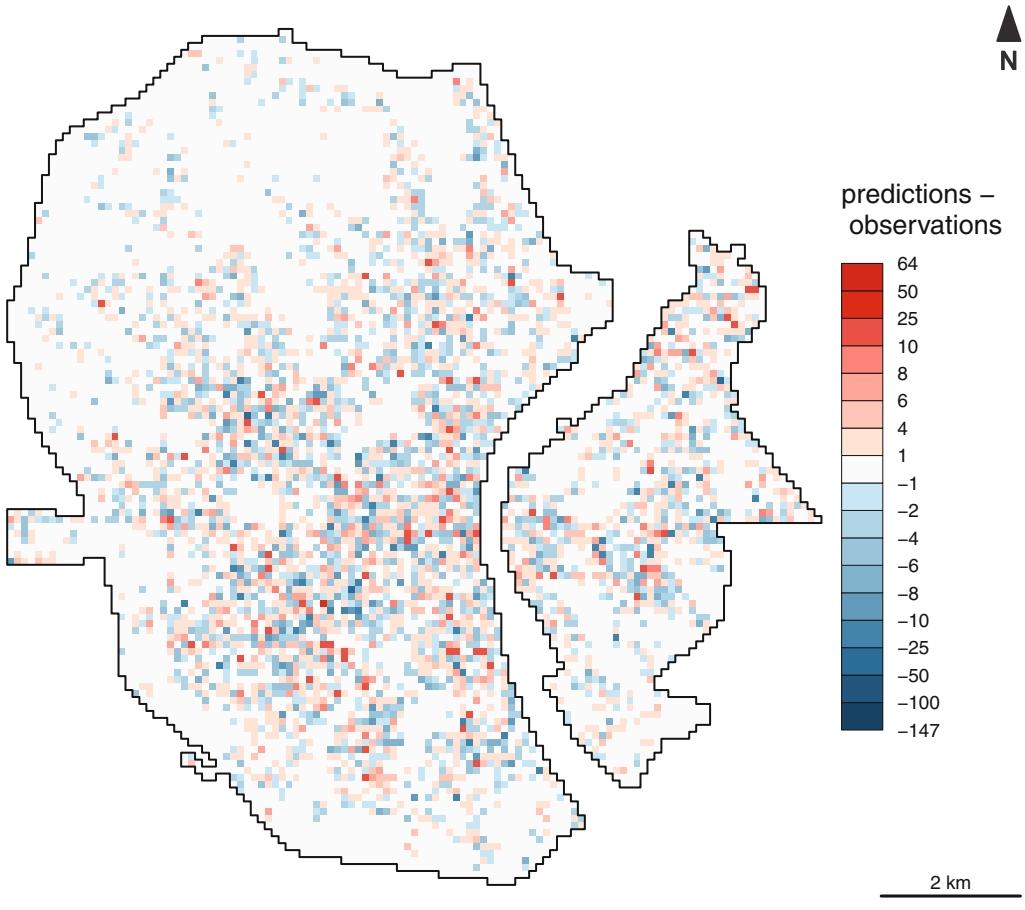

Figure 12. Difference between predicted and observed values of the test data set, predictions determined using XGBoost.

Flexibility

In comparison to linear regression or GLMs such as applied by (Tran and Ovtracht 2018; Gong and Yamamoto 2019; Bai and Jiao 2020; Huo et al. 2021), both methods allow to model the data much more flexibly, either using splines or regression trees. This allows to incorporate nonlinear relationships or interactions in a rather uncomplicated manner. But this advantage also leads to a higher complexity when defining a model to be estimated. In GAM, a model can be adapted through the selection of smoothing terms. Here, decisions have to be made regarding the variables to be smoothed, the type of smoothing spline and the dimension of the basis k for each smooth as well as the interactions between variables. In an XGBoost model, decisions have to be made for example regarding the hyperparameter optimization method, the parameters to be optimized and the ranges of values that are examined for these parameters. Regarding flexibility and ease of setting up a model, not one model can be clearly preferred over the other.

An advantage of XGBoost is its scalability allowing the flexible usage for different data and scenarios. While a different regression model structure has to be selected depending on the data set and distribution of the response variable, XGBoost allows to apply the same method

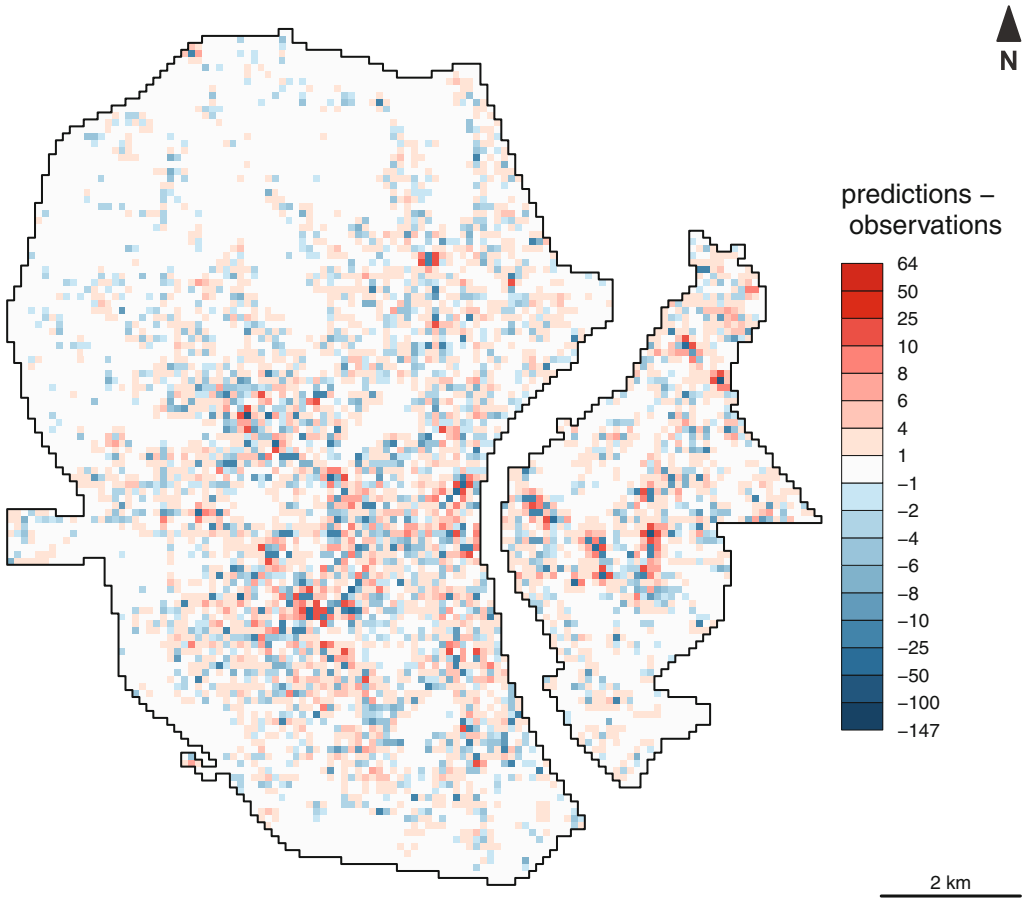

Figure 13. Difference between predicted and observed values of the test data set, predictions determined using geographically weighted XGBoost.

on a wide range of different data sets. Again, the tuning of hyperparameters allows to improve the model, but even without tuning, good predictions can be made. Of course, it has to be kept in mind that both methods are only applied on one data set in this analysis and the conclusions cannot be directly transferred on other data sets. But in general, XGBoost proved to generate good results also in previous studies working with various data sets.

Interpretation

A major disadvantage of machine learning methods in comparison to parametric regression methods is that the result cannot be interpreted in an equal way. Nevertheless, XGBoost offers a relatively straightforward interpretation method through importance plots. This allows to evaluate the importance of each variable in the decision-making process of the final model (Yang et al. 2020b). Still, the direction of the influence of a variable on the response variable is not directly accessible, thus hindering interpretation. In contrast to a parametric regression model, it is not possible to test hypotheses regarding the influence of variables. Therefore, the usage of XGBoost is somewhat limited when it comes to interpreting single variables.

Still, also a GAM cannot be interpreted as easily as a GLM as an effect of its flexibility in including smooth terms. Instead of a single estimation for each parameter, partial effect plots have to be evaluated that allow a deeper insight into the relationship of the variable's influence but complicate the interpretation. GAMs are an extension of GLMs that allow some of the flexibility of machine learning models while still being easier interpretable than such models, but more complicated than GLMs (Molnar 2019). Overall, regarding the possibilities of model interpretation, GAMs can be rated higher than XGBoost.

All in all, for both methods specific advantages and shortcomings can be found. While for a fixed set of tuning parameters the performance of XGBoost is superior in terms of computation time, a GAM allows to account for spatial interdependence and a more detailed interpretation. Regarding the focus of this analysis, the prediction of the response variable number of bike-sharing trips per grid cell, the performance of XGBoost can be rated higher. For both RMSE and sMAPE, XGBoost reaches lower values than all parametric regression models.

Discussion and conclusions

The preceding analysis dealt with the evaluation of parametric regression models in comparison to XGBoost for modeling a spatial data set consisting of the numbers of observed bike-sharing trips on the level of 100×100 m grid cells. For this purpose, several parametric regression models were implemented to model the influence of various spatial variables on the bike-sharing usage. Models considered were NB, Tweedie, zero-inflated NB regression and three GAM models with spatial splines: zero-inflated Poisson, NB and Tweedie. Among them, the Tweedie model with spatial splines led to the most consistent fit and predictions considering the AIC, sMAPE, and RMSE and was chosen for further interpretation and comparison to the XGBoost model. Both models were used to estimate predictions that were compared to the test data set through plots, RMSE and sMAPE.

The outcomes lead to the conclusion that XGBoost performs better regarding predictive accuracy. This result is in line with previous research comparing XGBoost to parametric approaches (Sathishkumar, Park, and Cho 2020; Ramesh et al. 2021). The predictions generated by the geographically weighted XGBoost model lead to RMSE = 3.72 and sMAPE = 0.24 in comparison to RMSE = 6.58 and sMAPE = 0.33 for the Tweedie GAM. Nevertheless, parametric regression models return additional information that help to interpret the results of the analysis and determine spatial influence factors on bike-sharing (Wang et al. 2015; Bao, Shi, and Zhang 2018; Ji et al. 2018).

Both models allow to handle spatial data and meet the requirements of spatial autocorrelation. In GAM, spatial coordinates can be directly included in the model estimation, whereas XGBoost does not offer this option. It has been applied to model temporal sharing data numerous times (Sathishkumar, Park, and Cho 2020; Yang et al. 2020b; Alencar et al. 2021), but in this study could be adapted to spatial data through the additional estimation of a GWR to account for spatial effects. The estimation of a GWR is a common technique to deal with spatial autocorrelation, but usually not in combination with XGBoost (Bao, Shi, and Zhang 2018; Ji et al. 2018; Wang et al. 2020b). This process also helped to smooth out the predictions, lowering variance between grid cells in spatial vicinity. The issue of overdispersion can be met by both models through the selection of adequate underlying distributions for the response variable. Geographically weighted NB models could be applied as well to deal with overdispersion (Da Silva and Rodrigues 2013; Chen et al. 2020).

In conclusion, the choice of modeling methods largely depends on the objective of the analysis: If it aims at the estimation of a model with high predictive power, XGBoost offers a fast and accurate approach. If the interpretation of influence factors is the focus of the research, GAMs include better options to assess the relationship of the dependent variable to various covariables while still offering flexible modeling methods and especially the option to include spatial data, and sufficiently high predictive power.

The models developed in the study can be applied to gain further insight into an existing bike-sharing system. As only publicly available data is used, the models can also be adopted to create estimates of bike-sharing usage in relation to spatial influence factors before the implementation of a bike-sharing system (Lee and Sener 2020).

A shortcoming of this analysis lies in the structure of the used data set that is not ideal as training and test data are not completely independent but the explanatory variables assume identical values for the corresponding observation in the training and test data set. Therefore, an optimal model regarding the training data set would perfectly predict the observations in the training data set which means the RMSE would effectively measure the difference between the observations in the training and test data. Still, this method of splitting the data in a training and test data set was chosen as a spatial split involves different problems. Therefore, RMSE and sMAPE have to be interpreted with caution as this method can lead to artificially low values. But as these measures are only used to compare models validated on the same data set, this should not influence the results in a manner that makes them invalid.

An option to increase interpretability of the XGBoost model would be to perform variable selection. Even though variables for each tree are selected automatically, the removal of those that are of very little importance for the modeling would lead to less complex models and importance plots. In GAM, several variables could already be removed from the models without compromising their predictive abilities, as certain variables assume only few different values or follow an extremely skewed distribution.

The implemented analysis could be extended to include additional modeling approaches such as Lasso-regularization, other tree-based methods such as random forest, or different algorithms or deep learning approaches. While this analysis was carried out on a case study of bike-sharing data, the findings of this study have the potential to be transferred to the modeling of other spatial distributions considering count data sets, such as the demand of different micromobility services.

Additionally, a spatiotemporal analysis could be performed using the underlying data set. This would allow to analyze temporal influence factors on bike-sharing usage and their possible interactions with the spatial ones considered in this analysis. Both methods, GAM and XGBoost, could generally be applied in such cases as well.

Acknowledgments

We would like to thank He Huang for his review and support regarding XGBoost. Open Access funding enabled and organized by Projekt DEAL.

References

Alencar, V. A., L. R. Pessamilio, F. Rooke, H. S. Bernardino, and A. Borges Vieira. (2021). "Forecasting the Carsharing Service Demand Using Uni and Multivariable Models." *Journal of Internet Services and Applications* 12(1), 1–20.

Armstrong, J. S. (1985). *Long-Range Forecasting: From Crystal Ball to Computer*, 2nd ed. New York: Wiley.

Bai, S., and J. Jiao. (2020). "Dockless E-Scooter Usage Patterns and Urban Built Environments: A Comparison Study of Austin, TX, and Minneapolis, MN." *Travel Behaviour and Society* 20, 264–72.

Bao, J., X. Shi, and H. Zhang. (2018). "Spatial Analysis of Bikeshare Ridership With Smart Card and POI Data Using Geographically Weighted Regression Method." *IEEE Access* 6, 76049–59.

Bivand, R. and D. Yu (2020). *spgwr: Geographically Weighted Regression*. R package version 0.6-34.https://CRAN.R-project.org/package=spgwr.

Botchkarev, A. (2018). Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. *arXiv preprint arXiv:1809.03006*.

Brunsdon, C., S. Fotheringham, and M. Charlton. (1998). "Geographically Weighted Regression." *Journal of the Royal Statistical Society: Series D (The Statistician)* 47(3), 431–43.

Caspi, O., M. J. Smart, and R. B. Noland. (2020). "Spatial Associations of Dockless Shared e-Scooter Usage." *Transportation Research Part D: Transport and Environment* 86, 102396.

Cepeda-Cuervo, E., M. Córdoba, and V. Núñez-Antón. (2018). "Conditional Overdispersed Models: Application to Count Area Data." *Statistical Methods in Medical Research* 27(10), 2964–88.

Chen, J., L. Liu, L. Xiao, C. Xu, and D. Long. (2020). "Integrative Analysis of Spatial Heterogeneity and Overdispersion of Crime with a Geographically Weighted Negative Binomial Model." *ISPRS International Journal of Geo-Information* 9(1), 60.

Chen, T. and C. Guestrin (2016). "XGBoost." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 10.1145/2939672.2939785

Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li (2021). *xgboost: Extreme Gradient Boosting*. R package version 1.3.2.1. https://CRAN.R-project.org/package=xgboost.

Cheng, J., X. Chen, J. Ye, and X. Shan. (2021). "Flow-Based Unit Is Better: Exploring Factors Affecting Mid-Term OD Demand of Station-Based One-Way Electric Carsharing." *Transportation Research Part D: Transport and Environment* 98, 102954.

Cheng, L., J. Yang, X. Chen, M. Cao, H. Zhou, and Y. Sun. (2020). "How Could the Station-Based Bike Sharing System and the Free-Floating Bike Sharing System Be Coordinated?" *Journal of Transport Geography* 89, 102896.

Correa, D., K. Xie, and K. Ozbay 2017). Exploring the Taxi and Uber Demand in New York City: An Empirical Analysis and Spatial Modeling. In *Proceedings of the 96th Annual Meeting of the Transportation Research Board*. Washington, DC.

Da Silva, A. R., and T. C. V. Rodrigues. (2013). "Geographically Weighted Negative Binomial Regression–Incorporating Overdispersion." *Statistics and Computing* 24(5), 769–83.

Dormann, C. F., J. M. McPherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl, R. G. Davies, A. Hirzel, W. Jetz, and D. W. Kissling. (2007). "Methods to Account for Spatial Autocorrelation in the Analysis of Species Distributional Data: A Review." *Ecography* 30(5), 609–28.

Dunn, P. K., and G. K. Smyth. (2005). "Series Evaluation of Tweedie Exponential Dispersion Model Densities." *Statistics and Computing* 15(4), 267–80.

Gebhart, K., and R. B. Noland. (2014). "The Impact of Weather Conditions on Bikeshare Trips in Washington, DC." *Transportation* 41, 1205–25.

Gong, L., and T. Yamamoto. (2019). "Temporal and Spatial Pattern of Shared Bike Trips - An Empirical Study of New York City." *Journal of the Eastern Asia Society for Transportation Studies* 13, 1333–47.

Guidon, S., H. Becker, and K. Axhausen (2019). "Avoiding Stranded Bicycles in Free-Floating Bicycle-Sharing Systems: Using Survival Analysis to Derive Operational Rules for Rebalancing." In *Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 1703–1708.

Hilbe, J. M. (2011). *Negative Binomial Regression*, 2nd ed. Cambridge: Cambridge University Press.

Hosseinzadeh, A., M. Algomaiah, R. Kluger, and Z. Li. (2021). "Spatial Analysis of Shared E-Scooter Trips." *Journal of Transport Geography* 92, 103016.

Hu, S., C. Xiong, Z. Liu, and L. Zhang. (2021). "Examining Spatiotemporal Changing Patterns of Bike-Sharing Usage During COVID-19 Pandemic." *Journal of Transport Geography* 91, 102997. https://www.sciencedirect.com/science/article/pii/S0966692321000508

Huang, H., M. Pouls, A. Meyer, and M. Pauly. (2020). "Travel Time Prediction Using Tree-Based Ensembles." In *Computational Logistics*, 412–27, edited by E. Lalla-Ruiz, M. Mes, and S. Voß. Cham: Springer International Publishing.

Huo, J., H. Yang, C. Li, R. Zheng, L. Yang, and Y. Wen. (2021). "Influence of the Built Environment on E-Scooter Sharing Ridership: A Tale of Five Cities." *Journal of Transport Geography* 93, 103084.

Ji, Y., X. Ma, M. Yang, Y. Jin, and L. Gao. (2018). "Exploring Spatially Varying Influences on Metro-Bikeshare Transfer: A Geographically Weighted Poisson Regression Approach." *Sustainability* 10(5), 1526.

Kaggle (2015). *Bike Sharing Demand: Forecast use of a City Bikeshare System*. San Francisco, CA: Kaggle. https://www.kaggle.com/c/bike-sharing-demand/overview/description

Kaviti, S., M. M. Venigalla, S. Zhu, K. Lucas, and S. Brodie. (2018). "Impact of Pricing and Transit Disruptions on Bikeshare Ridership and Revenue." *Transportation* 47(2), 641–62.

Lee, K., and I. N. Sener. (2020). "Emerging Data for Pedestrian and Bicycle Monitoring: Sources and Applications." *Transportation Research Interdisciplinary Perspectives* 4, 100095.

Lesmeister, C. (2017). *Mastering Machine Learning with R: Advanced Prediction, Algorithms, and Learning Methods with R 3.x*, 2nd. ed. Birmingham: Packt Publishing.

Li, A., and K. W. Axhausen. (2019). "Comparison of Short-Term Traffic Demand Prediction Methods for Transport Services." In *Arbeitsberichte Verkehrs- und Raumplanung*, 1–16. Zurich: IVT, ETH Zurich.

Li, A., P. Zhao, Y. Huang, K. Gao, and K. W. Axhausen. (2020). "An empirical Analysis of Dockless Bike-Sharing Utilization and Its Explanatory Factors: Case Study from Shanghai, China." *Journal of Transport Geography* 88, 102828.

Li, L. (2019). "Geographically Weighted Machine Learning and Downscaling for High-Resolution Spatiotemporal Estimations of Wind Speed." *Remote Sensing* 11(11), 1–26.

Liao, F., and G. Correia. (2020). "Electric Carsharing and Micromobility: A Literature Review on Their Usage Pattern, Demand, and Potential Impacts." *International Journal of Sustainable Transportation* 16(3), 269–86.

Liu, C., M. Zhao, W. Li, and A. Sharma. (2018a). "Multivariate Random Parameters Zero-Inflated Negative Binomial Regression for Analyzing Urban Midblock Crashes." *Analytic Methods in Accident Research* 17, 32–46.

Liu, Z., Y. Shen, and Y. Zhu. (2018b). *Inferring Dockless Shared Bike Distribution in New Cities: Technical Presentation*. CA, USA: Marina Del Rey.

Mardalena, S., P. Purhadi, J. D. T. Purnomo, and D. D. Prastyo. (2022). "The Geographically Weighted Multivariate Poisson Inverse Gaussian Regression Model and Its Applications." *Applied Sciences* 12(9), 4199.

Meyer, H., C. Reudenbach, S. Wöllauer, and T. Nauss. (2019). "Importance of Spatial Predictor Variable Selection in Machine Learning Applications–Moving from Data Reproduction to Spatial Prediction." *Ecological Modelling* 411, 108815.

Molnar, C. (2019). "Interpretable Machine Learning." In *A Guide for Making Black Box Models Explainable*. Munich: Christoph Molnar. https://christophm.github.io/interpretable-ml-book/

Perlman, J., and S. S. Roy. (2021). "Analysis of Human Movement in the Miami metropolitan Area utilizing Uber Movement Data." *Cities* 119, 103376.

Putatunda, S., and K. Rama. (2018). "A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost." In *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*, 6–10. Shanghai, China: Association for Computing Machinery. https://doi.org/10.1145/3297067.3297080

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Rabenstein, B. (2015). *Öffentliche Fahrradverleihsysteme - Wirkungen und Potenziale*, Vol 54. Veröffentlichungen aus dem Institut für Straßen- und Verkehrswesen: Universität Stuttgart.

Ramesh, A. A., S. P. Nagisetti, N. Sridhar, K. Avery, and D. Bein (2021). "Station-Level Demand Prediction for Bike-Sharing System." In *Proceedings of the 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 0916–0921*.

Reck, D. J., H. Haitao, S. Guidon, and K. W. Axhausen. (2021). "Explaining Shared Micromobility Usage, Competition and Mode Choice by Modelling Empirical Data from Zurich, Switzerland." *Transportation Research Part C: Emerging Technologies* 124, 102947.

Roy, A., T. A. Nelson, A. S. Fotheringham, and M. Winters. (2019). "Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists." *Urban Science* 3(2), 62.

Ryu, S.-E., D.-H. Shin, and K. Chung. (2020). "Prediction Model of Dementia Risk Based on XGBoost Using Derived Variable Extraction and Hyper Parameter Optimization." *IEEE Access* 8, 177708–20.

Sanders, R. L., A. Frackelton, S. Gardner, R. Schneider, and M. Hintze. (2017). "Ballpark Method for Estimating Pedestrian and Bicyclist Exposure in Seattle, Washington." *Transportation Research Record* 2605(1), 32–44.

Sathishkumar, V. E., J. Park, and Y. Cho. (2020). "Using Data Mining Techniques for Bike Sharing Demand Prediction in Metropolitan City." *Computer Communications* 153, 353–66.

Sawalha, Z., and T. Sayed. (2006). "Traffic Accident Modeling: Some Statistical Issues." *Canadian Journal of Civil Engineering* 33(9), 1115–24.

Köln, S. (2019). *Statistisches Jahrbuch Köln 2018*. Cologne: Office for Urban Development and Statistics.

Sun, Z., Y. Li, and Y. Zuo. (2019). "Optimizing the Location of Virtual Stations in Free-1003 Floating Bike-Sharing Systems with the User Demand during Morning and Evening Rush Hours." *Journal of Advanced Transportation* 2019, 1–11.

Tobler, W. R. (1970). "A computer movie simulating urban growth in the Detroit region." *Economic Geography* 46(sup1), 234–40.

Tran, T. D., and N. Ovtracht. (2018). "Promoting Sustainable Mobility by Modelling Bike Sharing Usage in Lyon." *IOP Conference Series: Earth and Environmental Science* 143, 1–12.

Wagner, S., T. Brandt, and D. Neumann. (2016). "In free float: Developing Business Analytics support for carsharing providers." *Omega* 59, 4–14.

Wang, N., J. Guo, X. Liu, and T. Fang. (2020a). "A Service Demand Forecasting Model for One-Way Electric Car-Sharing Systems Combining Long Short-Term Memory Networks with Granger Causality Test." *Journal of Cleaner Production* 244, 118812.

Wang, T., S. Hu, and Y. Jiang. (2021a). "Predicting Shared-Car Use and Examining Nonlinear Effects Using Gradient Boosting Regression Trees." *International Journal of Sustainable Transportation* 15(12), 893–907.

Wang, X., G. Lindsey, J. E. Schoner, and A. Harrison. (2015). "Modeling Bike Share Station Activity: The Effects of Nearby Businesses and Jobs on Trips to and from Stations." *Journal of Urban Planning and Development* 142, 4–15.

Wang, X., Z. Cheng, M. Trépanier, and L. Sun. (2021b). "Modeling Bike-Sharing Demand Using A Regression Model with Spatially Varying Coefficients." *Journal of Transport Geography* 93, 103059.

Wang, Z., L. Cheng, Y. Li, and Z. Li. (2020b). "Spatiotemporal Characteristics of Bike Sharing Usage around Rail Transit Stations: Evidence from Beijing, China." *Sustainability* 12(4), 1–19. https://www.mdpi.com/2071-1050/12/4/1299

Wollschläger, D. (2014). *Grundlagen der Datenanalyse mit R: Eine anwendungsorientierte Einführung*. überarb. und erw. Aufl. Statistik und ihre Anwendungen, Vol 3. Berlin: Springer Spektrum.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman & Hall / CRC Texts in Statistical Science, 2nd ed. Portland: CRC Press.

Yang, H., J. Huo, Y. Bao, X. Li, L. Yang, and C. R. Cherry. (2021). "Impact of E-Scooter Sharing on Bike Sharing in Chicago." *Transportation Research Part A: Policy and Practice* 154, 23–36.

Yang, H., Y. Zhang, L. Zhong, X. Zhang, and Z. Ling. (2020a). "Exploring Spatial Variation of Bike Sharing Trip Production and Attraction: A Study Based on Chicago's Divvy System." *Applied Geography* 115, 102130.

Yang, Y., A. Heppenstall, A. Turner, and A. Comber. (2020b). "Using Graph Structura information about Flows to Enhance Short-Term Demand Prediction in Bike-Sharing Systems." *Computers, Environment and Urban Systems* 83, 1–12.

Zhang, L., J. Cheng, and C. Jin. (2019). "Spatial Interaction Modeling of OD Flow Data: Comparing Geographically Weighted Negative Binomial Regression (GWNBR) and OLS (GWOLSR)." *International Journal of Geo-Information* 8, 1–18.

Zhang, W., R. Buehler, A. Broaddus, and T. Sweeney. (2021). "What Type of infrastructures do E-Scooter Riders Prefer? A Route Choice Model." *Transportation Research Part D: Transport and Environment* 94, 102761.

Zhao, D., G. P. Ong, W. Wang, and X. J. Hu. (2019). "Effect of Built Environment On Shared Bicycle Reallocation: A Case Study on Nanjing, China." *Transportation Research Part A* 128, 73–88.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web site.