Spatial modeling of bike-sharing trip data: A methodological comparison

Katja Schimohr*[1], Philipp Doebler[2], and Joachim Scheiner[3]

[1]PhD student, Department of Spatial Planning, Transport Research Group, TU Dortmund University, Germany

[2]Professor, Department of Statistics, Research Group of Statistical Methods in Social Sciences, TU Dortmund University, Germany

[3]Professor, Department of Spatial Planning, Transport Research Group, TU Dortmund University, Germany

SHORT SUMMARY

Usage data plays a major role in evaluating and planning sharing systems such as bike-sharing. Hence, effective methods are needed to analyze and model this kind of data. In this research, trip data of the bike-sharing system in Cologne, Germany is modeled. We compare two methods that can be applied in the modeling of spatial trip data, facing the requirements of spatial autocorrelation, zero-inflation and count data simultaneously. A generalized additive model (GAM) based on a Tweedie distribution is compared to a machine learning approach using the XGBoost algorithm. While the results of the GAM are easier to interpret and allow for the direct integration of spatial interdependencies in the model estimation, XGBoost leads to more precise predictions and can potentially be estimated in a shorter amount of time.

Keywords: bike-sharing, generalized additive model, machine learning, spatial data analysis, trip modeling

1. INTRODUCTION

Currently, more and more types of shared mobility options emerge in urban areas, among them car-, bike-, scooter- and e-scooter sharing. Ride-hailing services such as Uber serve similar functions. They supplement the existing urban transport system and increase users' flexibility in trip making by multimodal or intermodal mobility. All these options have in common that they are organized digitally. The generated usage data holds valuable information about urban mobility that can be used to better understand the usage of such sharing systems. The analysis of starting and ending points of trips allows to efficiently plan and manage sharing systems and to integrate them into the urban transport system.

In previous research, (zero-inflated) negative binomial regression has been found to be an adequate model to investigate sharing system usage (Bai & Jiao, 2020). Generalized additive models (GAM) underlying a negative binomial distribution of the dependent variable can be successfully applied as well (Hu, Xiong, Liu, & Zhang, 2021). The analysis of spatial datasets usually raises the question of spatial autocorrelation that requires further attention: it can be dealt with through the estimation of a geographically weighted regression (GWR) (Caspi, Smart, & Noland, 2020). Machine learning approaches are increasingly applied to model and predict the usage of sharing systems (Cheng, Chen, Ye, & Shan, 2021). A small number of studies apply XGBoost in the context of sharing systems (Yang, Heppenstall, Turner, & Comber, 2020). In contrast to

this research, most of these studies model the temporal distribution to predict demand over time. Comparisons between different machine learning approaches for modeling sharing data usually cannot determine one single best model, as different models perform best according to different evaluation metrics (A. Li & Axhausen, 2019). Still, XGBoost often ranks among the best models (Sathishkumar, Park, & Cho, 2020).

In this paper, a case study of the free-floating bike-sharing system of *nextbike* in Cologne, Germany, is conducted to compare modeling methods. Usage data was collected over a period of 5 weeks in 2019 and is used to calculate the number of bike-sharing trips that started within each cell of a 100 m grid laid over the operating area. This variable is to be modeled in relation to spatial influence factors both by a GAM and a machine learning approach using XGBoost. After training the models, predictions are generated that can be compared with a test dataset. To the best of our knowledge, no methodological research has yet addressed the specifics of modeling count data, an excess number of zeroes and spatial autocorrelation simultaneously. This paper aims at comparing two methods that meet these requirements. Besides predictive accuracy, ease of interpretation and computation time are evaluated.

2. METHODOLOGY

The underlying dataset contains the locations of all bikes in the *nextbike* system in Cologne that were collected from 30.09.2019, 10:46 am until 04.11.2019, 10:49. The data was scraped from `https://offenedaten-koeln.de` through an excel VBA script saving the locations every 15 minutes in a csv-table. All tables are combined and changes in location indicating the completion of a trip are extracted. All preprocessing and data analysis is performed in R. Trips that started or ended outside of Cologne or are shorter than 100 m are removed to exclude unlikely trips and repositionings of parked bikes. After preprocessing, 76,859 trips are included in the dataset. Then, the operation area of *nextbike* is divided into 8,955 grid cells of 100x100 m size and the number of trips that started within each of them is summarized over the whole observation period. These values are used as the dependent variable to be modeled.

Possible spatial influence factors on bike-sharing to be included in the model as independent variables are selected based on previous research. They include different land uses, points of interest (POI), the distance to universities and elements of the public transit system such as light rail and bus stations, population density and the share of age groups within each grid cell. All variables are presented in Table 1.

The bike-sharing usage data is split into a training and a test dataset that contain approximately the same number of trips using the 3,263 different times of observation. All trips are assigned subsequently according to their start time, alternating with every time of observation, until the second last time. After this process, there are 38,875 trips included in the training dataset and 37,965 trips in the test dataset. For both datasets, the number of trips that started through the whole observation period is summarized per grid cell.

Table 1: Description of all variables included in the dataset

Feature	Description	Unit
Trip count	Number of bike-sharing trips that started in each grid cell	Count
FNP 0	Residential building land	Percentage of grid cell
FNP 1	Special residential building land	
FNP 2	Mixed building land	
FNP 3	Commercial land	
FNP 4	Industrial land	
FNP 8	Water area	
FNP 9	Land for supply and disposal	
FNP 11	Land for community facilities	
FNP 12	Railway land	
FNP 13	Land for trains	
FNP 15	Core area	
FNP 16	Mixed area	
FNP 17	Special building land	
FNP 18	Special building land for a specific purpose	
FNP 21	Redevelopment area	
Green spaces	Green spaces of at least $100\,m^2$	Percentage of grid cell
Buildings	Buildings of all kinds	
Shops, food outlets, bars	Stores, service providers, restaurants, cafés, bars, nightclubs	Number of all POI within 3x3 grid cells
Healthcare facilities	Doctors, hospitals, retirement homes	
Schools	Elementary, secondary schools	
Kindergartens	Kindergartens	
Museums	Museums	
Event venues	Event venues, theatres, movie theatres, art exhibitions	
Libraries	Libraries	
Public institutions	Buildings of the city administration, Courthouses	
Sports facilities	Sports centers, gyms, soccer fields, running tracks	
Tourist attractions	Sights, tourist information	
Hotels	Hotels	
Places of worship	Places of worship of all kinds	
Playgrounds	Playgrounds	
University	Building of a university/ university of applied sciences	Distance to the closest object of the corresponding category in m/1,000 (1 if distance > 1,000 m)
Bus	Bus station	
Light rail	Light rail	
Arterial road	Arterial road	
Water	Water body of at least $10,000\,m^2$	
Population	Total population	Total number per grid cell
Shared flats	Total number of shared flats	
0-17 years	Share of persons aged 0-17 years	Share of persons of the respective age group living within 3x3 grid cells
18-29 years	Share of persons aged 18-29 years	
30-49 years	Share of persons aged 30-49 years	
50-64 years	Share of persons aged 50-64 years	
x-coordinate	x-coordinate	Coordinate system: ETRS89/UTM zone 32N
y-coordinate	y-coordinate	

Parametric model

As a parametric model, a GAM based on a Tweedie distribution is estimated that outperformed a negative binomial model in a preliminary analysis. This distribution family offers very flexible options to be adapted and allows to generalize many other distributions to model extremely skewed data (Dunn & Smyth, 2018, p.458). A GAM is a semi-parametric model that allows integrating smooth functions of variables as part of the predictor variables (Wood, 2017, p.161).

Smooth functions are included as penalized thin plate regression splines to deal with nonlinear effects of variables. Additionally, they enable variable selection and to control for spatial auto-correlation through the integration of a two-dimensional smooth term $f(x,y)$ based on the x- and y-coordinates of each grid cell's centroid. The GAM is fit by penalized iteratively re-weighted least squares (PIRLS) including a penalty for the wiggliness of each smooth function (Wood, 2017, p.249). Smoothing parameters are estimated using a Laplace approximation for restricted maximum likelihood (REML).

XGBoost

The XGBoost algorithm 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 that enables merging several weak learners into a stronger model. Among a broad range of modeling options, XGBoost can be applied for Poisson regression and Tweedie regression. As a decision tree on its own can suffer from overfitting, measures to lower the variance are taken, such as regularization, shrinkage and column subsampling (Chen & Guestrin, 2016).

While it is possible to estimate XGBoost models without hyperparameter tuning using the default values, the adjustment leads to much more appropriate models (Ryu, Shin, & Chung, 2020). There are over 30 hyperparameters in XGBoost, but it is sufficient to tune only a smaller share. As general model settings, `booster = "gbtree"`, `eta` = 0.3, `eval_metric = "rmse"` in combination with `maximize = FALSE` and `early_stopping_rounds` = 10 are chosen.

The optimal number for `nrounds` is determined through cross-validation using the default values for the XGBoost model. The parameters `max_depth, gamma, subsample`, `colsample_bytree` and `min_child_weight` are optimized in a second step applying a grid search approach. For the parameters to be optimized, the following values are considered, mostly based on (Huang, Pouls, Meyer, & Pauly, 2020) and requiring the estimation of $11,000$ models:

- `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` $\in \{0,10,20,30,40,50,60,70,80,90,100\}$

XGBoost itself does not account for spatial interdependence. Therefore, a GWR is calculated on the geo-referenced predictions made by an XGBoost model to create smoother predictions, following the approach of (L. Li, 2019). The predictions determined by the GWR are the final predictions of the geographically weighted XGBoost.

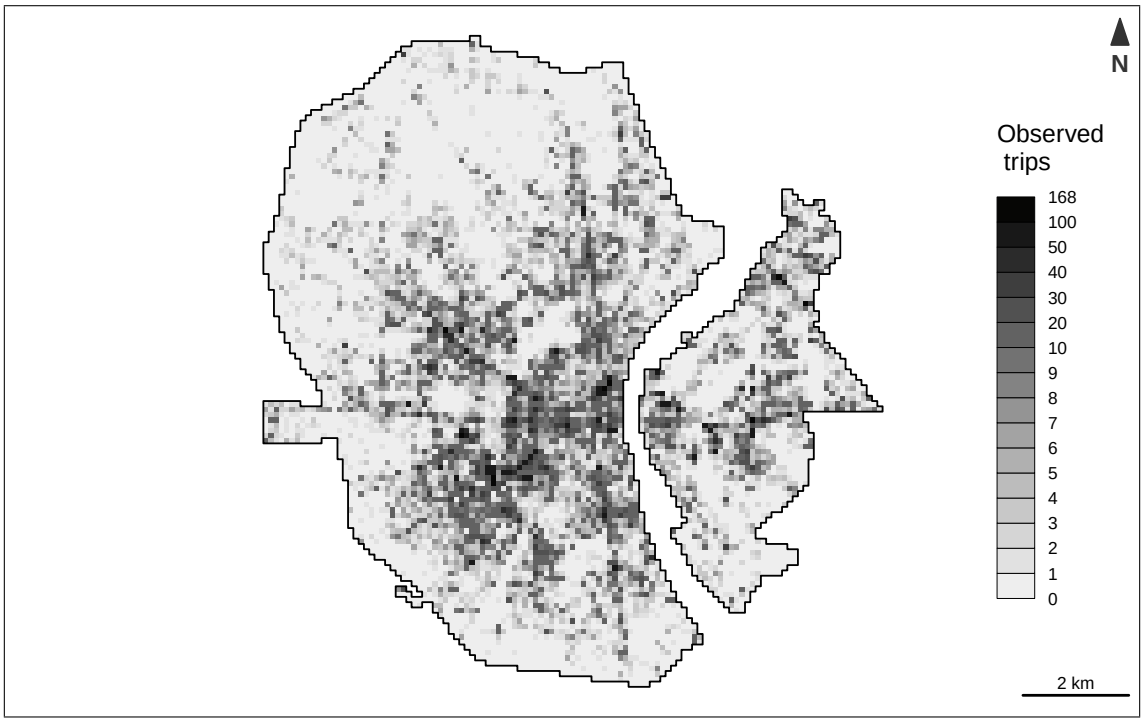

Figure 1: Spatial distribution of the number of trips per grid cell in the test dataset.

3. RESULTS AND DISCUSSION

On average, 8.52 trips per grid cell could be observed. There are several upper extreme values and the distribution of observed values is right-skewed: The maximum number is 329 and for 3,218 of 8,955 grid cells, 0 trips were observed. In Figure 1, the spatial distribution of trips in the test dataset is displayed. Trips appear to be clustered in certain parts of the study area, especially in the city center, while cells with no trips seem to concentrate on the outer areas. The independent variables chosen in this study (see Table 1) exhibit sufficiently low correlations. Most of them lie between -0.2 and 0.2, a majority even between -0.1 and 0.1.

Parametric model

In the following, the results of the modeling using a Tweedie GAM are presented. To choose an adequate value for the number of basis dimensions k for each variable, multiple models are estimated starting with $k = 3$. k is successively increased until sufficient. Then, variables that are insignificant in the model, have low effective degrees of freedom (edf) values < 1 and/or where a partial influence plot shows a straight line at 0, are excluded from the model. After this process, the following variables were removed: *FNP 1, FNP 4, FNP 9, FNP 11, FNP 16, FNP 18, FNP 21, healthcare 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*.

The parameters estimated for the selected variables in the Tweedie GAM are specified in Table 2. For smooth terms, edf indicating the complexity of a smooth and test statistics used in an ANOVA test to test the significance of the smooth are displayed. The reference degrees of freedom (ref.df) used in computing test statistics represents $k - 1$. F and the corresponding p-value represent the results of the ANOVA test.

The predictions generated by the Tweedie GAM are displayed in Figure 2. The structure seems to be similar to Figure 1, indicating a good model performance.

Table 2: Estimate, Std. error, 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	Std.error	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

XGBoost

The two objectives `count:poisson` and `reg:tweedie` are both included in the tuning process to determine the one leading to a better fit. As the Poisson model leads to much smaller values of the average root mean square error (RMSE) in the grid search, this distribution is chosen. Tuning is aimed at minimizing the RMSE for the training data, resulting in RMSE = 2.97 and symmetric mean absolute percentage error (sMAPE) = 0.19 for the test data. This process leads to the following parameters for the Poisson model: `max_depth` = 33, `gamma` = 0, `subsample` = 0.9, `colsample_bytree` = 0.5, `min_child_weight` = 0, `nrounds` = 57. Additionally, 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.

In the next step, a GWR is estimated using the predictions generated in XGBoost as input values, leading to a bandwidth of 51.45 [m]. The predictions generated by geographically weighted XGBoost lead to a higher RMSE = 3.72 and sMAPE = 0.24. Their spatial structure is displayed in Figure 3. The geographically weighted XGBoost leads to smoother predictions that appear comparable to the predictions generated by the Tweedie regression. Therefore, the more natural fit of a geographically weighted XGBoost comes at the expense of a decrease in the accuracy of predictions.

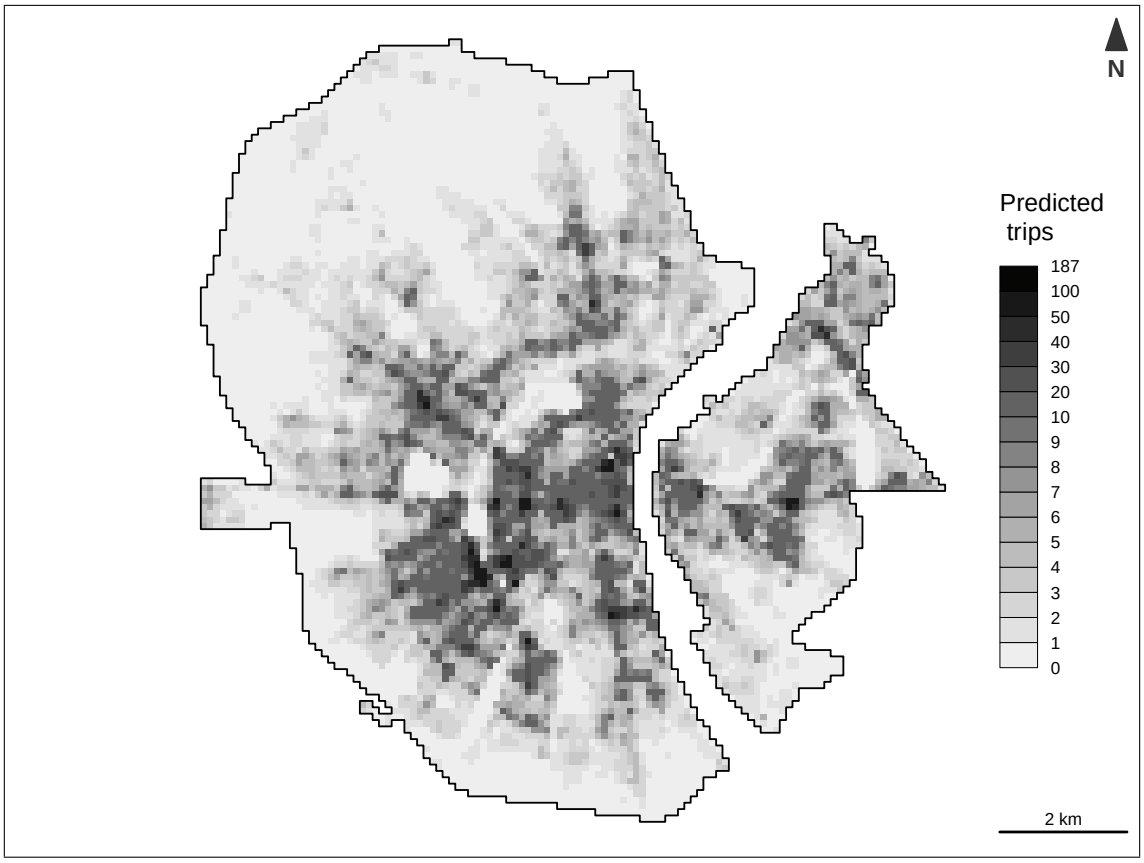

Figure 2: Predictions of bike-sharing trip counts determined by a Tweedie GAM.

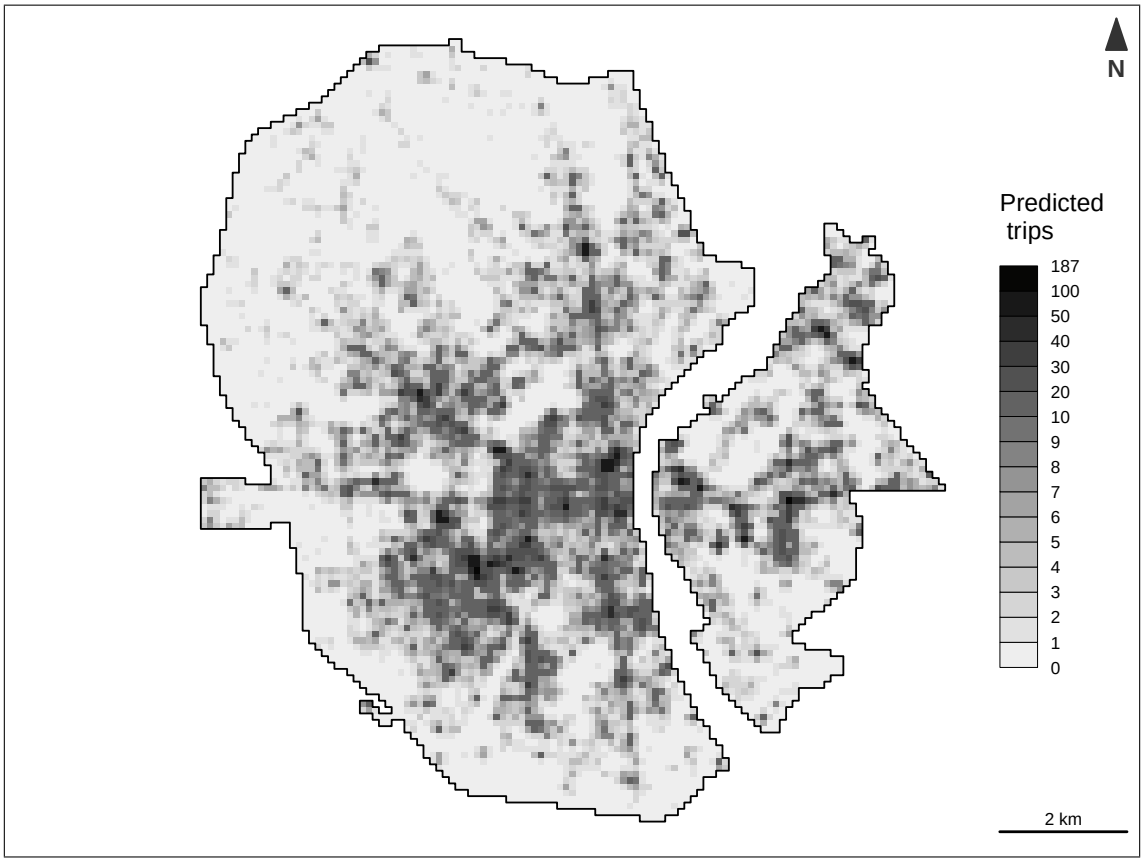

Figure 3: Predictions of bike-sharing trip counts determined by geographically weighted XGBoost.

Table 3: RMSE, sMAPE and computation time.

	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
Hyperparameter tuning	-	3:23:32.56 h	-

Comparison

Both XGBoost models lead to much smaller values than the parametric approach (see Table 3). In addition, the estimation of a single XGBoost model takes far less time than the estimation of a GAM with the selected dataset and settings. Still, it has to be considered that for an optimal model fit, tuning is necessary when estimating an XGBoost model requiring much more time.

Both methods allow to flexibly adapt to the dataset and to incorporate non-linear relationships or interactions, either using splines or regression trees. This advantage comes at the cost of a more complex model definition. Therefore, both modeling approaches' flexibility can be rated similarly.

Regarding interpretability, parametric methods are to be preferred as the modeling output usually gives a clear indication of each parameter's influence. Still, in a GAM, additional means such as partial effect plots are required to interpret the relationship between the dependent and the explanatory variables. The same applies to XGBoost. Here, importance plots allow an interpretation regarding the strength of each variable's influence on the decision-making process of the model. Still, this method does not allow to determine the direction of the relationship and to perform hypothesis tests. In comparison, the interpretation of XGBoost is somewhat limited.

An advantage of XGBoost over parametric models is its scalability which allows the application of one model for a wide range of datasets. In contrast, the choice of an adequate parametric model structure is necessary when using a parametric model to adapt to specific datasets and distributions of the response variable. As this research only deals with the modeling of one dataset, this is not further demonstrated but can be deducted from the model structure of XGBoost and its application in previous studies.

In this research, especially the ability of each model to deal with spatial autocorrelation is of interest. The interaction between the x- and y-coordinate is included in the GAM through a spline modeling an interaction term. This allows to interpret the relation of the dependent variable to the locations in space. In XGBoost, x- and y-coordinates are included separately as regular variables. To incorporate spatial relationships, an additional GWR is performed. The option to directly include spatial information in the model estimation is an advantage of GAMs in comparison to XGBoost. Additionally, if coordinates are included in XGBoost, it is crucial to adopt settings that prevent the model to fit exclusively to the coordinates.

In total, both methods entail specific qualities that set them apart from the other but no method can be clearly preferred. Instead, models should be chosen depending on the priorities of the research. XGBoost succeeds at estimating extremely accurate predictions and should be applied whenever this is the aim of the research. Regarding computation time, XGBoost allows to create good predictions in an extremely short amount of time when omitting the hyperparameter tuning process. When there is a focus on model interpretation, GAMs offer considerably more information. Additionally, GAMs offer a better way to deal with the requirements of spatial autocorrelation.

4. CONCLUSIONS

In this analysis, a GAM based on a Tweedie distribution and a geographically weighted XGBoost were applied on a spatial dataset containing the numbers of observed bike-sharing trips within 100x100 m grid cells.

XGBoost results in lower values of the RMSE and sMAPE. This leads to the conclusion that, only regarding predictive accuracy, XGBoost should be preferred. Nevertheless, parametric models entail certain advantages that should be considered: the results offer an easier interpretation and spatial coordinates can be directly included in the model estimation.

As the explanatory variables assume the same values both in the training and in the test dataset, the resulting RMSE and sMAPE may be too low and should not be compared to other studies. Still, both methods were applied on the same dataset which means that the comparison of methods within this study is still valid.

In future research, variable selection could be performed to create smaller models. Also, an extension to include other methods, especially machine learning approaches could broaden the scope of this research.

ACKNOWLEDGMENT

We would like to thank He Huang for his support regarding XGBoost.

REFERENCES

Bai, S., & Jiao, J. (2020). Dockless e-scooter usage patterns and urban built environments: A comparison study of austin, tx, and minneapolis, mn. *Travel Behaviour and Society*, *20*, 264–272. doi: 10.1016/j.tbs.2020.04.005

Caspi, O., Smart, M. J., & Noland, R. B. (2020). Spatial associations of dockless shared e-scooter usage. *Transportation Research Part D: Transport and Environment*, *86*, 102396. doi: 10.1016/j.trd.2020.102396

Chen, T., & Guestrin, C. (2016, 8). Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Retrieved from http://dx.doi.org/10.1145/2939672.2939785 doi: 10.1145/2939672.2939785

Cheng, J., Chen, X., Ye, J., & Shan, X. (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. doi: 10.1016/j.trd.2021.102954

Dunn, P. K., & Smyth, G. K. (2018). *Generalized linear models with examples in r*. New York, NY: Springer New York. doi: 10.1007/978-1-4419-0118-7

Hu, S., Xiong, C., Liu, Z., & Zhang, L. (2021). Examining spatiotemporal changing patterns of bike-sharing usage during covid-19 pandemic. *Journal of Transport Geography*, *91*, 102997. Retrieved from https://www.sciencedirect.com/science/article/pii/S0966692321000508 doi: https://doi.org/10.1016/j.jtrangeo.2021.102997

Huang, H., Pouls, M., Meyer, A., & Pauly, M. (2020). Travel time prediction using tree-based ensembles. In E. Lalla-Ruiz, M. Mes, & S. Voß (Eds.), *Computational*

logistics (pp. 412–427). Cham: Springer International Publishing.

Li, A., & Axhausen, K. W. (2019). Comparison of short-term traffic demand prediction methods for transport services. *Arbeitsberichte Verkehrs- und Raumplanung*, 1–16. doi: 10.3929/ethz-b-000356143

Li, L. (2019). Geographically weighted machine learning and downscaling for high-resolution spatiotemporal estimations of wind speed. *Remote Sensing*, *11*(11), 1-26. doi: 10.3390/rs11111378

Ryu, S.-E., Shin, D.-H., & Chung, K. (2020). Prediction model of dementia risk based on xgboost using derived variable extraction and hyper parameter optimization. *IEEE Access*, *8*, 177708-177720. doi: 10.1109/ACCESS.2020.3025553

Sathishkumar, V. E., Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. *Computer Communications*, *153*, 353–366. doi: 10.1016/j.comcom.2020.02.007

Wood, S. N. (2017). *Generalized additive models: An introduction with r, second edition*. Portland: CRC Press.

Yang, Y., Heppenstall, A., Turner, A., & Comber, A. (2020). Using graph structural information about flows to enhance short-term demand prediction in bike-sharing systems. *Computers, Environment and Urban Systems*, *83*, 1–12. doi: 10.1016/j.compenvurbsys.2020.101521