Estimating bicycle counts with Strava data along a gradient of use intensities

1. Abstract

Reliable information on daily bicycle traffic provides a fundamental basis for city planners and scientists. We apply Generalised Boosted Regression Models to estimate daily bicycle counts for different German locations with different degrees of urbanisation. Altogether 44,136 daily datapoints from 46 counter locations covering a time period of four years were considered. Crowdsourced fitness tracker data from Strava, socio-demographics, land use and weather data were used as independent variables. Our results indicate that weather has the strongest influence on estimated bicycle counts, exceeding the relevance of fitness tracker data. In an overall model daily bicycle counts were estimated with a mean absolute percentage error (MAPE) of 27.9%. In terms of location specific estimations, a MAPE of 11.2% was reached. With our approach, high-quality out-of-sample predictions are also feasible. Based on our estimations we assume the volatility of fitness tracker user share to have a major impact on model accuracy.

Keywords:
daily bicycle traffic
bicycle counts
spatial differences
crowdsourced data
strava
calibration

2. Introduction

As studies show cycling has become increasingly popular over the last decades (Pucher & Buehler, 2017), probably because it is health-promoting and helps reduce ones personal carbon footprint. Increasing numbers of cyclists make it important for city and traffic planners to comprehend which factors affect spatial and and temporal cycling activities. The research project AQUATAG, within which this study is embedded, aims to understand spatial distribution and dynamics of recreational activities in (e.g., swimming, diving), on (e.g. paddling, sailing) and along (e.g. cycling, walking) surface waters to assess the effects on aquatic ecosystems and potential conflicts between different activity types. In case of cycling, comprehensive data sets of bicycle counts are missing, which makes it more difficult to study the drivers affecting cycling activity. Traditionally, bicycle count could only be obtained from on-site monitoring, which limited the number and duration of observations. During the last years automated bicycle counters became more and more common, but the number of such automated counters is still small and they are often placed at larger and highly frequented streets. The increasingly common use of fitness tracker apps, such as STRAVA, added new possibilities to estimate bicycle counts or bicycle traffic volumes (both

expressions are used synonymously) for unmonitored areas. However, a potential hurdle in the use of such crowdsourced data is the biased and rather small user group. Accordingly, different studies tried to assess the representativity of these data (Lee & Sener, 2021) and furthermore use it as a basis for actual bicycle count estimation, recently with help of machine learning approaches (Miah et al., 2022).
With the present study we aim to build a generalisable model for the estimation of bicycle traffic counts to better understand the cyclist impacting dynamics and the spatial differences. We do this by using STRAVA data, representing one possible source of crowdsourced data, and monitored count data as a basis together with weather data, socio-demographics and land use data. Further, in the context of a project addressing water-based recreation, our study, for the first time, also uses distance to surface water bodies as a potential driver. Our data set comprises 46 locations covering inner city areas, outskirts, and less urban areas in and around larger cities in Germany, allowing us to identify the major drivers explaining temporal and spatial differences in bicycle counts and to test the accuracy of our models under different site conditions. Last but not least, we approach the question of how model accuracy is affected by the share of bicyclists providing crowdsourced data. The outcomes are assumed to contribute additional knowledge for traffic related planners as well as recreational management.

3. Background and research questions

3.1. Literature review

The literature provides a variety of different approaches to estimate bicycle counts. Dadashova and Griffin (2020) mention scaling methods, direct-demand modelling and time series amongst others as rather traditional methods. As such Hankey et al. (2012) and Hankey et al. (2017) provide scaling methods and modelling approaches for estimating bicycle traffic based on weather data, nearby built environment and socio-demographics, as well as street characteristics. During the last years it became more popular to use data from crowdsourced fitness tracker apps, with Strava leading the way. Livingston et al. (2021) predict bicycle traffic volumes for the city of Glasgow including regression models for so called out-of-sample predictions in order to estimate numbers for subsequent years. Nelson et al. (2021) choose input data from different cities to develop a more generalisable model. Lee and Sener (2021) provide a literature review on the use of Strava Metro data for bicycle monitoring. Besides pointing out the opportunities of Strava data, they also stress potential challenges when using this data (e.g. “under-representativeness of the general population, bias towards and away from certain groups, and lack of demographic and trip details at the individual level”). This is in line with Conrow et al. (2018), Griffin and Jiao (2019) and Watkins et al. (2016) who report an overrepresentation of males and age groups between 25 and 44 years for example. Studies like Roy et al. (2019) try to correct biases in Strava data by using Poisson regression including surrounding land use and socio-demographics. However, previous studies have also shown that linear models including Strava data are susceptible to high errors because of non-linear relationships between monitored count and Strava count data (Miah et al., 2021). Recent approaches take advantage of machine learning techniques

combined with fitness tracker data besides other explanatory variables (see for example Al-Ramini et al. (2022), Kamalapuram (2022) or Miah et al. (2022)). Miah et al. (2022) tested a variety of different machine learning approaches to predict bicycle traffic for the city of Portland. According to their random forest model results, Strava count is ranked as the most relevant driver amongst all explanatory variables, whereas their XGBoost model ranks speed limit as the most important driver followed by Strava count.

3.2. Research and knowledge gaps

In spite of a recently growing body of bicycle count prediction studies, the overall number of such publications is still rather low and remaining knowledge gaps could be identified. As Jestico et al. (2016) state, prediction models may only be suitable for locations with similar conditions. Thus, robust models should include input data from heterogenious and diverse locations, which was not always taken into account in previous studies.

The examination of the relative influence of the considered predictors was rather approached in recent studies (cf. Miah et al. (2022), Al-Ramini et al. (2022) and Kamalapuram (2022)) with partly ambiguous results concerning their ranking, indicating a need for further research.

However, the potential to predict bicycle counts for unknown locations and time periods which have not been included in training data sets at all is yet fairly unstudied and hardly addressed in the available literature. Besides that, former studies mostly only considered locations with medium to high bicycle traffic, ignoring areas representing lower bicycle counts. Furthermore, Dadashova and Griffin (2020) state that future research should approach how the share of bicyclists using fitness tracker apps affects the accuracy of prediction models which are based on these data. For our analysis we include a wide range of different bicycle count locations covering different use intensities but also different shares of bicyclists using the fitness tracker Strava, in order to generate representative results ensuring additional value for the scientific literature.

3.3. Research questions

Based on our literature review we form the following research questions:

I. *Which are the most important predictors to estimate bicycle counts and how does their relative influence differ?*
II. *How accurate are daily bicycle count estimations for unknown locations and time spans?*
III. *How is model accuracy affected by the share of bicyclists providing crowdsourced data for estimation models?*

4. Methods and Data

4.1. Input Data

Hourly biking counts of the years 2017-2020 were provided by Strava_Inc. (2020) for the Spree-Havel and the Ruhr catchments, both representing study river basins of the AQUATAG project. Monitored bicycle count data for locations within the scope of the project areas were enquired from local administration authorities. Altogether, hourly and daily monitored bicycle count data have been available for 46 locations within the three German states Berlin

(17), Brandenburg (14) and North Rhine-Westphalia (NRW) (15) (cf. Figure 1) for a time period of four years (2017-2020), including information from permanent and temporary counting stations.

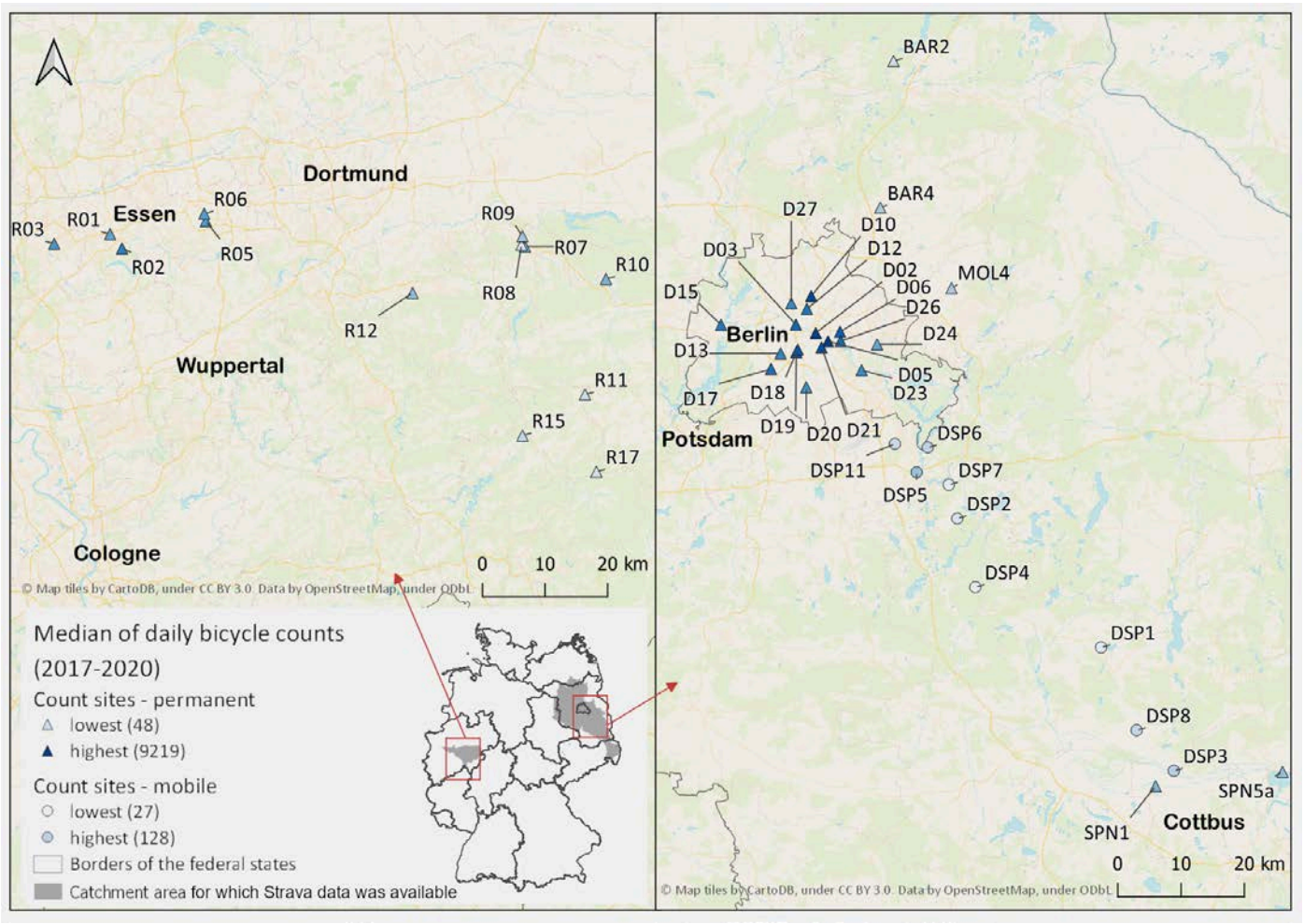

Figure 1: Locations of the bicycle counters used in the study.

Each bicycle counting location was assigned to one or more OpenStreetMap (OSM) edges, since Strava data were provided in these spatial units. Thus, corresponding bicycle counts and STRAVA user counts could be derived.

For each counter location the mean daily air temperature (in C°) and the daily precipitation sum (in mm) were extracted from the E-OBS dataset with a spatial resolution of 0.1° (Cornes, 2018). Furthermore, we derived the absolute population as well as the population density per square kilometre from the latest available official census data, provided as grid with a cell size of 100 m * 100 m (Statistisches_Bundesamt, 2011).
In similar studies age, gender and income are often the most commonly used socio-demographic predictors. While income data was not available in an adequate spatial resolution, spatial information on age and gender could be received for December 2020 at administrative district level (mean size of 78.5 km² considering all available districts in all three states) from the statistical offices of Berlin-Brandenburg (Berlin-Brandenburg, 2021) and North Rhine-Westphalia (Nordrhein-Westfalen, 2022) and was assigned to the residential areas per district. Land use information on residential areas (Corine Land Cover Class “Urban fabric”) were taken from a landcover map LBM by (BKG, 2018).

In order to substitute the lacking income data, spatial data on the composition of milieus was used. Here Sinus-Milieus® data were used, provided as a two-dimensional classification of citizens according to their social situation (lower class, middle class or upper class) and basic orientation (tradition, individualisation and reorientation) using age, gender, education and income to describe the different Sinus-Milieus® groups (Flaig & Barth, 2014). In our study the following groups are considered: expeditious milieu, hedonistic milieu, conservative established milieu, liberal intellectual milieu, performer milieu, adaptive pragmatic milieu, precarious milieu, socio-ecological milieu and traditional milieu.

To derive milieu information for different buffer zones around the bicycle counter locations, we calculated weighted means under the consideration of population density (census data), residential area distribution and the respective demographic information at postal code level. Furthermore, we used the following features from OpenStreetMap (2022) to derive the bike path length within a given buffer zone: all key features of the class "highway" with the assigned value "cycleway", all key features of the class "cycleway", all features of the class "bicycle" tagged as "yes" and all features of the class "bicycle" tagged as "designated".

Information on land use categories and levels of soil sealing were obtained from BKG (2018) to calculate average soil sealing levels and the relative share of the major land use types. According to Bossard et al. (2000) we distinguished “urban fabric” (CLC-code: 111, 112), “industrial, commercial and transport units” (CLC-code: 121, 122, 123, 124), “mine, dump and construction sites” (CLC-code: 131, 132, 133), “green urban areas” (CLC-code: 141, 142), “agricultural areas” (CLC-code: 211, 212, 213, 221, 222, 223, 231, 241, 242), “forest and near natural areas” (CLC-code: 311, 312, 313, 322, 323, 324, 331, 332, 333, 334, 335), “natural grassland” (CLC-code: 321), “wetlands” (CLC-code: 411, 412, 421, 422, 423) and “waterbodies” (CLC-code: 511, 512, 521, 522, 523). Additionally, the distance from each bicycle counting location to the nearest waterbody was calculated.

Information on all socio-demographic, land cover and land use data variables (cf. Table 1) was derived for each counter location for buffer widths of 250m, 500m, 1000m, 2000m, 4000m, 6000m, 8000m and 10000m.

To capture temporal bicycle traffic dynamics, we included the variables weekend, public holiday and school holidays to the analysis. As the number of STRAVA users increased considerably during the study period year and month were introduced as variables, also accounting for temporal effects.

Table 1: Initial set of independent variables for prediction models.

Category	Variable		Source
Temporal Variables	Year		Own calculations
	Month		Own calculations
	Weekend		Own calculations
	Public Holiday		Own calculations
	School Holiday		Own calculations
	Weekend, public holiday or school holiday		Own calculations
Strava Variables	Number of Strava user counts		Strava
	Percentage of commuting Strava users (%)		Strava
Weather Variables	Precipitation (mm)		E-OBS
	Average daily temperature (C°)		E-OBS
Land use, land cover and infrastructure	Sealing (%)		BKG
	Urban fabric (%)		BKG
	Industrial, commercial and transport units (%)		BKG
	Mine, dump and construction sites (%)		BKG
	Green urban areas (%)		BKG
	Agricultural areas (%)		BKG
	Forest and near natural areas (%)		BKG
	Natural grassland (%)		BKG
	Wetlands (%)		BKG
	Waterbodies (%)		BKG
	Distance to waterbodies (m)		BKG, manual
	Bike path length (m)		OSM
Socio-demographics	Absolute population [an]		Census
	Population density (people per km²)		Census
	Expeditious milieu [an]		Sinus, Census
	Hedonistic milieu [an]		Sinus, Census
	Conservative established milieu [an]		Sinus, Census
	Liberal intellectual milieu [an]		Sinus, Census
	Performer milieu [an]		Sinus, Census
	Adaptive pragmatic milieu [an]		Sinus, Census
	Precarious milieu [an]		Sinus, Census
	Socio-ecological milieu [an]		Sinus, Census
	Traditional milieu [an]		Sinus, Census
	Male, Age 0-5 [an]	Female, Age 0-5 [an]	Statistical Office
	Male, Age 5-10 [an]	Female, Age 5-10 [an]	Statistical Office
	Male, Age 10-15 [an]	Female, Age 10-15 [an]	Statistical Office
	Male, Age 15-20 [an]	Female, Age 15-20 [an]	Statistical Office
	Male, Age 20-25 [an]	Female, Age 20-25 [an]	Statistical Office
	Male, Age 25-30 [an]	Female, Age 25-30 [an]	Statistical Office
	Male, Age 30-35 [an]	Female, Age 30-35 [an]	Statistical Office
	Male, Age 35-40 [an]	Female, Age 35-40 [an]	Statistical Office
	Male, Age 40-45 [an]	Female, Age 40-45 [an]	Statistical Office
	Male, Age 45-50 [an]	Female, Age 45-50 [an]	Statistical Office
	Male, Age 50-55 [an]	Female, Age 50-55 [an]	Statistical Office
	Male, Age 55-60 [an]	Female, Age 55-60 [an]	Statistical Office
	Male, Age 60-65 [an]	Female, Age 60-65 [an]	Statistical Office
	Male, Age 65-70 [an]	Female, Age 65-70 [an]	Statistical Office
	Male, Age 70-75 [an]	Female, Age 70-75 [an]	Statistical Office
	Male, Age 75-80 [an]	Female, Age 75-80 [an]	Statistical Office
	Male, Age 80-85 [an]	Female, Age 80-85 [an]	Statistical Office
	Male, Age 85-90 [an]	Female, Age 85-90 [an]	Statistical Office
	Male, Age 90 upwards [an]	Female, Age 90 upwards [an]	Statistical Office

[an] *Absolute number*

To distinguish whether counter locations are based in rather urban or rather rural areas we invoke the classification of Eurostat (2022) to differentiate NUTS3 regions according to the following types: “predominantly rural”, “intermediate” and “predominantly urban”. Based on

the counter location within a certain NUTS3 region, one of the above stated settlements types is assigned according to Eurostat-Data (2022).
The bicycle count locations D02 (Berlin – Jannowitzbrücke), R03 (North Rhine-Westphalia – Mühlheim) and BAR2 (Brandenburg – Eichhorst) were selected as case examples, since they differ regarding their daily bicycle counts and each represents one of the three involved states of this study (cf. Figure 1).

4.2. Generalised Boosted (Regression) Models

We use “Generalised Boosted (Regression) Models” (GBM), also referred to as “Boosted Regression Trees” (BRT), since this method can be used to approach all three research questions. GBM combine decision tree models with the method of boosting, that is building an ensemble of many decision trees whereby every new decision tree is based on the remaining residuals between monitored and estimated bycicle counts of its predecessor. The models are not only able to estimate a dependent variable based on a set of independent variables, but furthermore state the relative influence of the latter. The relative influence of each variable can be calculated with the R-package “gbm” package. In accordance with Friedman (2001) the relative influence of a variable is based on how often on average, across all trees, it is chosen to split a decision tree. For a more detailed explanation of BRT and GBM see Ridgeway (2007) and Elith et al. (2008). Within GBMs, different settings of higher-level properties, so called hyperparameters, are possible. These hyperparameters have an effect on the complexity, the learning speed and consequently the model results (Chicco, 2017) and therefore the best performing settings may be identified by comparing model accuracy results based on different hyperparameter combinations. The fitting of hyperparameters requires a training of the model under consideration of different hyperparameter settings and can result in an excessive amount of computation time. For the present model set-up 517 independent variables, comprising all possible buffer widths, per 44,136 daily data points (cf. 5.1) would have to be considered for the training. The identification of the different variable contributes was therefore done using standard hyperparameter settings (see “Buffer Detection Width Model” in Table 6). The hyperparameter fitting was done subsequently only using selected high-influence variables, defined as the most important buffer width per variable, and after an exclusion of variables with a non-significant relative influence (< 0.1 %).

The hyperparameter tuning was done by performing a grid search in which the generalised boosted regression models were trained several times using different combinations of the following hyperparameter values: shrinkage rate (learning rate), interaction depth (maximum number of splits in each tree), minimum number of observations in trees' terminal nodes (“n.minobsinnode”). This results in 27 initial combinations of hyperparameter values.

The root-mean-square error (RMSE) between monitored and estimated bicycle counts was used as performance criterium. Hyperparameter combinations with the least RMSE were chosen as reference point for a second grid search step. After this we tested whether a recommended default bag fraction of 0.5 (Ridgeway, 2007) or a greater one of 0.65 yielded better results. Resulting hyperparameter values were used accordingly for further modelling. For the final models we calculated RMSE, R^2, the mean absolute percentage error (MAPE), the Nash–Sutcliffe model efficiency coefficient (NSE) and the percent bias (PBIAS) in order to compare their performance with reference to Table 2. Altogether we test six different

models (one standard model, one valuation model, one unknown year model and three different unknown location models). For the standard model the data set is randomly split into a training data set (70%) and a testing data set (30%). The validation model, however, used the entire data set as training data, which was also used to describe the overall model performance. In order to gain the model performance for each location, only data of the respective count location was used as training data. To determine the model performance for an untrained year (unknown year model), the training data set was restricted to a time span between 2017 and 2019, while 2020 was used for validation. Here again, we distinguish between an overall performance and performance for each location. For the unknown location models, in accordance to the other set-ups, data from one of the three locations was excluded from the training data sets and 2020 data was used for validation.

Table 2: Goodness of fit taken from Pérez-Sánchez et al. (2017) based on Moriasi et al. (2007).

	NSE	PBIAS
Very good	0.75 - 1.00	< ±10
Good	0.65 - 0.75	±10 - ±15
Satisfactory	0.50 - 0.65	±15 - 25
Unsatisfactory	< 0.50	± 25

To draw conclusions about which factors the MAPE depends on, we use linear regression models, estimating the MAPE of the standard model for each location based on the explanatory variables (i) median daily count, (ii) Strava user share, (iii) the standard deviation of Strava user share (as indicator for volatility), (iv) the number of daily datapoints per location which were included in the training data set and (v) the share of missing Strava values (Strava-NA) indicating the share of datapoints for which Strava information was missing.

5. Results

5.1. Descriptive statistics

Bicycle counts from May 30th until November 1st 2019 of location D05 (Berlin/Oberbaumbrücke) were excluded as construction works at the site led to a change of the traffic guidance (Hering, 2019), which resulted in significantly less monitored bicycle counts but hardly unchanged Strava counts. Excluding these, we altogether used 44,136 daily data points from 46 counter locations as input data for our models. Counting locations in Berlin show the largest variability but also the highest number in daily bicycle volumes (Figure 2) with the lowest median of 501 for D24 (Alberichstraße) based in the outskirts and the highest median of 9219 for D05 (Oberbaumbrücke) in the city centre. The median daily bicycle counts of the North Rhine-Westphalia (NRW) locations in the Ruhr catchment area range from 48 at R08 (Jahnallee in Arnsberg) to 1474 at R02 (Grugatrasse in Essen). The lowest medians were observed for the counting stations in Brandenburg reaching from 27 at DSP4 (B179 in Löpten) to 286 at SPN1 (Ringchaussee in Burg).

Using the settlement structure classification of Eurostat (2022) all counter locations in Berlin were classified as predominantly urban. All Brandenburg counter locations were assigned as intermediate. The western North Rhine-Westphalia counter locations (R01, R02, R03, R05, R06, R12) were classified as predominantly urban whereas the remaining eastern locations (R07, R08, R09, R10, R11, R13, R15, R16, R17) were classified as intermediate (cf. Figure 1).

Figure 3 compares the daily monitored counts in 2020 versus the Strava bicylce counts for the three locations D02, R03 and BAR2. All three locations show a similar seasonal pattern with higher counts during the summer months and lower counts in winter. In case of BAR2 Strava counts are almost consistently at 0 during the spring months. Strava user shares (calculated as the ratio between actual bicycle counts and Strava bicycle counts) for all locations grouped by state are shown in Figure 4. Average shares range between about 1% and approximately 4%. The Berlin locations show the lowest Strava user shares (mean = 0.87; median = 0.73), followed by the North Rhine-Westphalia based locations (mean = 3.81; median = 2.77) and the Brandenburg counter locations with the highest Strava user shares (mean = 5.47; median = 4.04). The maximum values reach up to 24% for the Berlin stations, up to 83% for North Rhine-Westphalia locations and in the case of Brandenburg 91%.

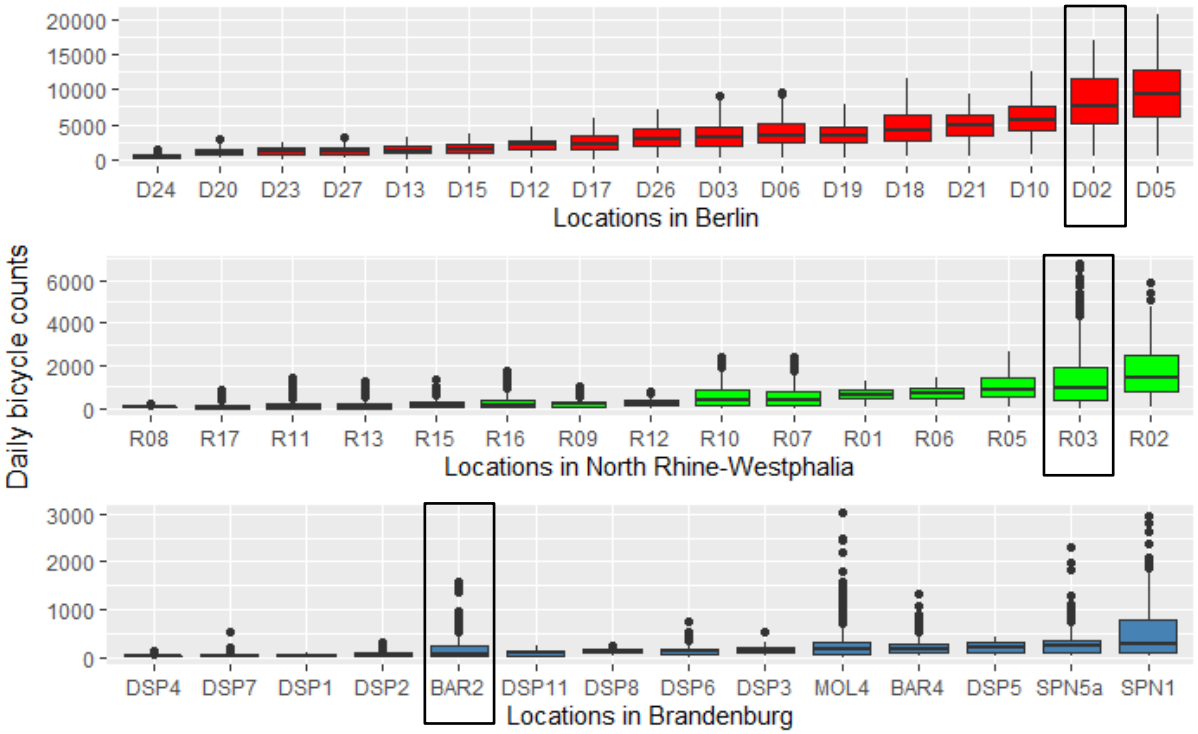

Figure 2: Daily bicycle counts for all counter locations grouped by federal state.

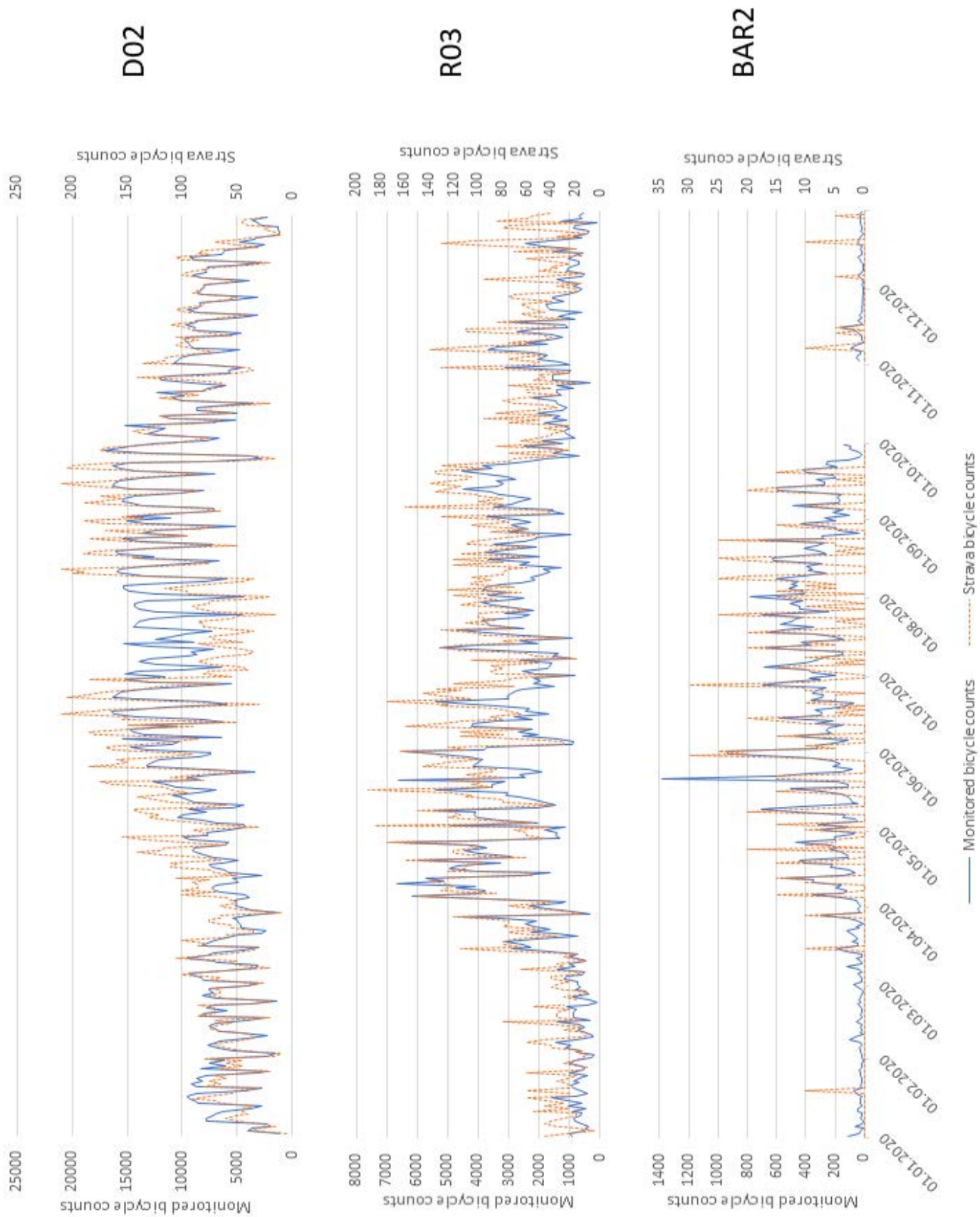

Figure 3: Comparison of monitored bicycle counts and Strava bicycle counts for the locations D02, R03 and BAR2.

**Bicycle counts for October 2020 were not available for location BAR2. Hence, this period was not included.*

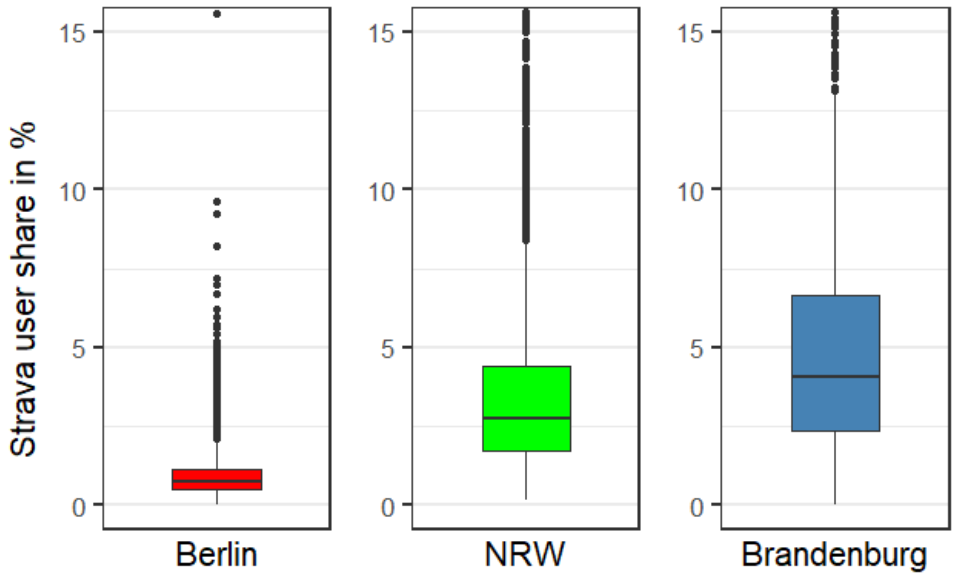

Figure 4: Daily share of Strava users in relation to daily bicycle counts per location in the respective federal state.

5.2. Variable selection and relative influence on dependent variable

The initial set of considered predictors consisted of 517 independent variables. From this a sub-set of 33 explanatory variables contributing the highest influence to estimated daily bicycle counts were selected for the further modelling tasks (Table 3). For visualised graphs of the average model estimation for varying values of the top three predictors in our standard model see Figure 6in the Appendix.

Table 3: Independent variables included in the final models and their relative influence in the fitted standard model for the estimation of daily bicycle counts.

ID	Variable (measuring unit)	Relative influence on daily bicycle volumes (%)
1	Average Daily Temperature (degree Celsius)	24.45
2	Percentage of commuting Strava users (%)	19.07
3	Precipitation (mm)	11.12
4	Male, Age 0-5 within 1000m (absolute number)	7.83
5	Number of Strava bicycle counts	7.01
6	Male, Age 25-30 within 500m (absolute number)	4.04
7	Month	3.06
8	Female, Age 40-45 within 4000m (absolute number)	2.04
9	Forest and Near-Natural Areas within 500m (%)	1.71
10	Year	1.70
11	Sealing within 250m (%)	1.50
12	Bike Path Length within 250m (m)	1.29

13	Distance to Water (m)	1.20
14	Waterbody within 10000m (%)	1.20
15	Industrial, Commercial or Transport Units within 250m (%)	1.18
16	Mine, Dump or Construction Site within 10000m (%)	1.17
17	Weekend, Public Holiday or School Holiday	1.14
18	Agricultural Area within 6000m (%)	1.08
19	Weekend	1.00
20	Urban Green within 500m (%)	0.95
21	Natural Grassland within 8000m (%)	0.81
22	Male, Age 30-35 within 2000m (absolute number)	0.71
23	Conservative Established Milieu within 4000m (absolute number)	0.68
24	Male, Age 90 upwards within 1000m (absolute number)	0.63
25	Urban Fabric within 250m (%)	0.60
26	Liberal Intellectual Milieu within 500m (absolute number)	0.44
27	Wetlands within 4000m (%)	0.42
28	Public Holiday	0.40
29	School Holiday	0.40
30	Female, Age 90 upwards within 4000m (absolute number)	0.37
31	Male, Age 85-90 within 4000m (absolute number)	0.35
32	Bourgeois Middle Class Milieu within 500m (absolute number)	0.24
33	Expeditious Milieu within 4000m (absolute number)	0.22

5.3. Model performance

Table 4 collates the statistical performance parameter of the different GBM subdivided into an overall performance for all locations and such for the three case example locations.

For all tested model variants, the results for Jannowitzbrücke (D02) with the second highest median bicycle count of all stations always provided the best statistical performance indicators, which, on the other hand were worst for Eichhorst (BAR2), having the lowest median counts out of the three case example locations.

Table 4: Model performance of different GBM for all locations (overall) and for three exemplary locations.

Model	All locations	Case study locations		
		Jannowitzbrücke (D02)	**Radschnellweg Mühlheim (R03)**	**Eichhorst (BAR2)**
Mean monitored annual bicycle counts	21,946,387	2,958,945	495,945	52,340
Strava user share in %	1.91 %	0.80 %	3.65 %	2.73 %
Validation (100 % data for training)	MAPE: 19.1 % R^2: 0.98 PBIAS: -0.4 % NSE: 0.98 RMSE: 377.2	MAPE: 9.3 % R^2: 0.95 PBIAS: -0.4 % NSE: 0.95 RMSE: 869.1	MAPE: 22.4 % R^2: 0.92 PBIAS: -1.9 % NSE: 0.92 RMSE: 334.1	MAPE: 26.3 % R^2: 0.87 PBIAS: -4.7 % NSE: 0.86 RMSE: 65.8
Standard (70 % data for training)	MAPE: 27.9 % R^2: 0.97 PBIAS: -0.5 % NSE: 0.97 RMSE: 433.2	MAPE: 13.3 % R^2: 0.91 PBIAS: 0.6 % NSE: 0.91 RMSE: 1178.8	MAPE: 29.2 % R^2: 0.89 PBIAS: -1.1 % NSE: 0.89 RMSE: 417.3	MAPE: 43.0 % R^2: 0.84 PBIAS: -2.0 NSE: 0.84 RMSE: 70.8
Unknown year (years 2017-2019 for training)	MAPE: 36.2 % R^2: 0.91 PBIAS: -10.6 NSE: 0.9 RMSE: 809.8	MAPE: 17.7 % R^2: 0.8 PBIAS: -0.7 % NSE: 0.8 RMSE: 1739.2	MAPE: 27.1 % R^2: 0.79 PBIAS: 0.3 % NSE: 0.77 RMSE: 657.1	MAPE: 50.7 % R^2: 0.78 PBIAS: -15.9 % NSE: 0.76 RMSE: 102.0
Unknown location (all data except case study location for training)		MAPE: 38.7 % R^2: 0.83 PBIAS: 24.6 % NSE: 0.53 RMSE: 2670.6	MAPE: 37.2 % R^2: 0.81 PBIAS: -38.8 % NSE: 0.34 RMSE: 1125.0	MAPE: 3253.0 % R^2: 0.35 PBIAS: 921.5 % NSE: -85.84 RMSE: 1927.2

Considering Table 2, our results indicate that the validation model has performed best which makes sense since here all data has been used for training, so that the indicators basically describe the performance of the model training and not of the validation. The standard model performed second best, followed by the unknown year models. The least good performance was observed for the unknown location models, where unsatisfying performance values were obtained for all three locations (PBIAS), with BAR2 standing out with the only NSE below 0. In all cases estimations for the location D02 (Jannowitzbrücke) outperformed estimations for the location R03 (Mühlheim) which in turn outperformed estimations for the location BAR2 (Eichhorst). Exemplary visualisations for the unknown year model can be found in Figure 5 the Appendix.

The upper graphs in Figure 7 display the median daily counts sorted in ascending order per location versus the MAPE (a) and versus the Strava user share (b), indicating higher median

daily counts are associated with lower Strava user shares and related to better model accuracy. The lower graphs show the standard deviation of Strava users sorted in ascending order versus the Strava user share (c) and versus the MAPE for the respective location (d), suggesting a relation of a higher Strava user volatility with an increased Strava user share and lower model accuracy.

We used linear regression models to identify potential drivers for MAPE derived for the individual monitoring locations. Table 5 shows the results for the estimation of the MAPE of the standard model based on several independent variables characterising each location (see Table 7 in the Appendix for the data basis). Those were the share of Strava counts relative to the actual bicycle counts (Strava user share) as well as the standard deviation of this Strava user share as an indicator for volatility, the absolute number of datapoints per location in the training data set, the share of missing Strava information (Strava-NA) relative to the number of datapoints and the median daily count. For the locations SPN1, SPN5a and DSP5 no more than one Strava count were available. Hence, these counter locations could not be included in the regression models.

Table 5: Linear regression results: Parameter estimates for MAPE estimation of the standard model and standard deviation in brackets.

	Model 1	Model 2	Model 3
(Intercept)	24.73 **	25.58 *	14.79
	(8.19)	(9.47)	(10.15)
Standard deviation of Strava user share	3.35 ***	3.32 ***	1.44
	(0.81)	(0.84)	(1.14)
Number of datapoints	-0.01 *	-0.01 *	-0.01
	(0.01)	(0.01)	(0.01)
Share of Strava-NA	0.35 **	0.34 **	0.31 *
	(0.11)	(0.12)	(0.12)
Median daily count	-	-0.00	-0.00
	-	(0.00)	(0.00)
Strava user share	-	-	3.17 *
	-	-	(1.38)
N	43	43	43
R2	0.70	0.70	0.74
	Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’		

6. Discussion

6.1. Descriptive statistics

Bicycle counters in urban regions show higher bicycle counts than such based in intermediate urban/rural regions. Unfortunately, according to Eurostat (2022), our input data does not feature any predominantly rural counter locations. However, eleven locations (R11, R17, BAR2, SPN1, DSP1, DSP2, DSP3, DSP4, DSP5, DSP7 and DSP8) fulfil the first of three OECD requirements of less than 300 inhabitants per km² as rural condition, when derived for a 1km buffer.

When comparing monitored bicycle counts and Strava counts for the same period in case of the predominantly urban counter location D02 (see Figure 3), both variables seem to show congruent temporal patterns at the beginning and at the end of 2020. A divergence can be observed around spring which can be attributed to an increase of Strava user share at the start of the corona pandemic and a decrease around school summer holidays. The scaling factor of 100:1 is in line with an average Strava user share of 1% for this location and the corresponding year.

For the less urban Brandenburg based location BAR2 (Eichhorst) with less bicycle counts and less Strava counts, the temporal patterns match not as good, as Figure 3 shows. In particular, the lower values are often underestimated as Strava counts less than 3 are set to 0 and higher counts are given in 5-count steps in order to ensure data privacy (Lee & Sener, 2021). In addition, the average share of Strava users is higher than at most Berlin locations but also varies much stronger over time. This variation could also be explained by the before mentioned rounding effect of using 5-count steps for Strava information, affecting smaller ratios between monitored counts and Strava counts much stronger as it is the case in more rural areas. Another interesting observation that can be made when comparing BAR2 and D02 are the inverted peaks. At D02 count peaks are found during the week, while at BAR2 these were observed during weekends. This is in line with the higher number of commuting cyclists in urban areas during the week as well as with the higher numbers of recreational cyclists in less urban areas on the weekends.

On average, Strava users shares decrease with the degree of urbanization, consequently being lowest in Berlin and highest but more volatile in Brandenburg. The Strava user shares of all North Rhine-Westphalia (NRW) locations, featuring both, urban and intermediate urban/rural areas, range between Berlin and Brandenburg. This pattern can be attributed to the fact that in order to reach a Strava user share below 1 %, a monitored count of at least 500 is needed, since the lowest Strava count greater than 0 is 5. In our dataset all Brandenburg locations have a median daily count below 500, whereas all Berlin locations show a median daily count above 500. In case of the North Rhine-Westphalia sites some counter locations are above and some are below this threshold.

6.2. Variable selection and relative influence on independent variable

The three dominating variables temperature, share of commuting stave users and precipitation together contribute 55 % of influence on the bicycle count prediction (Table 3). Weather related variables support the obvious assumption that at higher temperatures and no/less rain bicycle activities increase (cf. Figure 6). However, our model predictions indicate a certain temperature optimum around 22 °C indicating decreasing bicycle counts on hot days. In our model higher amounts of daily rain sums are associated with lower numbers of predicted bicycle counts.

Nevertheless, it can be observed that this relationship is not linear, showing an increase of estimated bicycle counts for increasing daily precipitation sums from around 25 mm until ca. 40 mm. We suppose that this is due to the fact that in both areas, North Rhine-Westphalia and Berlin-Brandenburg, the greatest precipitation amounts are observed in the warmer summer months, although this pattern is more pronounced in the latter case (ClimateData, 2022a, 2022b).While weather data explain temporal dynamics on a macro scale, the share of commuting Strava users contribute to explain temporal differences on a microscale (work day - weekend / holidays) and parts of the spatial differences between urban and more rural areas, as described above. As can be seen in Figure 6 in our model an increasing share of commuting Strava users leads to higher numbers of estimated bicycle counts.

From the selected 33 variables, 25 subordinated variables have an individual share of influence of ≤ 2 %, however, contributing 21.4 % in total. While the dominating variable are sufficient to describe the general spatio-temporal count dynamics, subordinated variables are supposed to explain local specifications at individual periods or locations. This would explain, why subordinated variables (e.g. month [7]), containing potentially correlated information of dominating variables (e.g. temperature [1]), do not have even lower influence. Similarly, four subordinated variables [17, 19, 28, 29] relate to workday-holiday information, which differ in the three considered states and contribute to explain these spatial differences. Another example are variables addressing land use [9, 11, 12, 13, 14, 15, 16, 18, 20, 21, 25, 27] with a total influence of 13.1% in our standard model. Urban green and presence of surface water were reported to be generally preferred by recreationist and a factor affecting well-being (Venohr et al., 2018) and might have a positive effect on the surrounding microclimate of a counter location. This relatively low share of influence, might be explained by several reasons: a) cyclists (in particular commuter) prefer other aspects (e.g. shorter distance, quality of asphalt, car density or other security aspects) over land-use attributed, b) available counters do not cover the combination of high bicycle counts in e.g. a forested area, c) the used climate date do not have the required spatial resolution and precision to display microclimate conditions.

6.3. Model performance and uncertainty

Our standard model yielded an overall MAPE of 27.9% and showed better average MAPE for the urban Berlin based counter locations (13.7%) compared to the partly urban and partly intermediate urban-rural NRW counter locations (45.4%) and the entirely intermediate urban-rural counter locations based in Brandenburg (56.2%). Our out-of-sample prediction models show good and useful prediction results for D02 but also less good results for other locations (cf. unknown location model for BAR2), with the unknown year models performing better than the unknown location models in all cases.

Dadashova and Griffin (2020) provide an overview of studies predicting bicycle activity including prediction accuracy where El Esawey et al. (2015) with 10.4% reported the lowest MAPE values for their Vancouver based prediction model, being not too far from our average MAPE for the Berlin based counter locations. The average MAPE of the corresponding study Dadashova and Griffin (2020) was 29% which is comparable to our standard model overall MAPE. In contrast to our results Jestico et al. (2016) observed higher model error for high volume sites compared to low volume sites (cf. Miah et al. (2022)).

Figure 7 shows that in our case counter locations with higher daily bicycle counts, most notably the Berlin locations, showed also decreased MAPE and lower Strava user shares but also a lower volatility of these user shares. According to Figure 7 d) locations with a higher MAPE show a higher volatility of Strava user shares, yet it has to be considered that higher Strava user shares are correlated with a higher volatility of Strava user shares (a Pearson test revealed a positive correlation of 0.8 between the two variables.). Therefore, excluding this variable from linear regression, as done in the linear regression model 1 and model 2 (Table 5) seemed sensible. Based on the statistically significant estimates of model 1 and model 2 it can be assumed that the MAPE for bicycle count predictions for a certain counter location depends on the standard deviation of Strava user shares as an indicator for volatility, the absolute number of datapoints in the training data set and the share of missing Strava information, which may be related to lower Strava counts being assigned as NA-values. Linear regression models showed that the daily count median did not have a significant effect on the MAPE (cf. model 2). While Jestico et al. (2016) did observe higher model errors for locations with increased bicycle counts our results show the opposite, i.e. increased errors for low bicycle counts . As a possible explanation we suggest elevated and highly volatile Strava user shares for low bicycle counts. This can arise from the rounding Strava counts in increments of five, having a bigger impact on smaller proportions (see Table 8 and

Table 9 in the Appendix). This in total leads to a blur of the Strava user shares and impairs the model accuracy at lower counts. Berlin stations account for the largest number of datapoints, the highest bicycle counts (with the least volatile user shares) and the least shares of missing Strava values (cf. Table 7), which leads to the best model performance for these stations.

7. Conclusion

In this study we have presented a machine learning approach that is able to estimate bicycle counts for urban and more rural locations, based on open source fitness tracker data, weather data, socio-demographics and land use information. Our prediction models suggest that even if bicycle count data from fitness tracker apps are available, weather variables and information on the bicyclist composition (share of commuters) reach a higher relative influence as predictors than the mere fitness tracker counts. A constant or uniformly changing Strava user share would allow estimating monitored bicycle counts by a simple factor or function. Consequently the volatility of fitness tracker user shares, can be seen as a main cause for model uncertainties and increased MAPE. Spatial differences in model accuracy are assumed to be based on this (being caused through rounding effects due to data privacy measures) as well as on the number of datapoints and the share of missing Strava data per counter location. Hence, estimation models based on Strava information should be used with caution for less frequented locations. As the daily mean temperature was found to be the main predictor of bicycle counts with increasing predicted counts until 22 °C and decreasing count predictions above this temperature, further research should focus on how the microclimate and associated land use features (e.g. blue-green infrastructure) affect bicycle count estimation.

8. Appendix

Table 6: Hyperparameters chosen after three step grid search.

Hyperparameter	Buffer Width Detection Model	Standard Model	Validation Model	Unknown Year Model	Unknown Location Models		
					D02	R03	BAR2
Shrinkage Rate	0.3	0.08	0.1	0.05	0.1	0.1	0.1
Interaction Depth	5	16	18	22	5	15	15
Minimum Number of Observations in Trees' Terminal Nodes	5	13	9	17	5	5	5
Bag Fraction	0.5	0.5	0.5	0.5	0.65	0.5	0.5
Number of Trees	5000	5000	5000	5000	5000	5000	5000
Cross Validations to Perform	10	10	10	10	10	10	10

Table 7: Data basis for linear regression of standard model MAPE estimation.

Station	Region	MAPE (standard model)	Mean Strava user share (%)	Median daily count	Standard deviation of Strava user share	Number of datapoints	Share of Strava-NA
D02	Berlin	13.29	0.80	7661	0.31	1461	6.16
D03	Berlin	13.32	0.96	3228	0.50	1461	8.42
D17	Berlin	17.03	1.30	2276	0.75	1461	11.98
D20	Berlin	13.12	1.23	1020	0.65	1461	24.71
D23	Berlin	16.08	0.64	1122	0.46	1453	49.14
D13	Berlin	13.98	0.61	1272	0.39	1461	44.35
D15	Berlin	13.16	0.72	1493	0.61	1461	35.73
D05	Berlin	12.48	0.65	9219	1.25	880	13.98
D18	Berlin	14.65	0.68	4219	0.37	1461	11.16
D24	Berlin	18.51	1.09	501	1.10	1461	64.34
D26	Berlin	13.79	1.24	3039	0.56	1461	10.47
D27	Berlin	11.19	1.15	1125	0.60	1461	27.58
D19	Berlin	11.47	0.64	3490	0.31	1461	10.27
D21	Berlin	11.86	0.46	4783	0.22	1461	10.75
D10	Berlin	13.37	0.62	5566	0.46	1461	8.76
D12	Berlin	11.36	0.84	2097	0.47	1461	14.92
D06	Berlin	13.57	1.15	3395	0.52	1461	10.34
R13	NRW	83.53	6.04	90	9.73	1435	57.42
R12	NRW	45.44	4.39	237.5	3.17	838	39.38
R11	NRW	105.17	5.04	80	6.27	1159	69.28
R17	NRW	59.64	4.18	67	13.08	1279	81.24
R15	NRW	48.18	3.46	123.5	3.36	1344	70.83
R03	NRW	29.2	3.65	957.5	2.19	1096	9.12

R07	NRW	32.34	3.01	420	2.86	972	30.66
R02	NRW	18.39	3.71	1474	1.85	945	2.86
R01	NRW	20.59	0.83	638.5	0.40	938	56.72
R16	NRW	57.45	8.00	128	9.07	752	38.83
R05	NRW	30.79	2.36	880	1.63	365	7.40
R06	NRW	19.57	1.47	686	0.96	365	28.22
R10	NRW	45.01	1.27	398	1.01	619	59.13
R08	NRW	50.62	7.47	48	3.91	93	83.87
R09	NRW	34.74	5.31	199	4.24	93	25.81
BAR4	Brandenburg	28.09	5.99	178	5.99	1395	56.63
SPN1	Brandenburg	32.07	*NA*	286	*NA*	1429	100.00
BAR2	Brandenburg	43.04	2.73	78.5	2.93	1358	76.66
MOL4	Brandenburg	33.81	5.41	175	5.19	1372	38.63
SPN5a	Brandenburg	25.15	*NA*	231	*NA*	1355	100.00
DSP1	Brandenburg	102.53	12.71	45.5	4.12	76	97.37
DSP7	Brandenburg	86.75	16.17	39	15.82	104	75.96
DSP11	Brandenburg	85.94	7.96	90	5.95	72	43.06
DSP5	Brandenburg	27.03	4.24	223	*NA*	59	98.31
DSP4	Brandenburg	91.47	11.07	27	5.97	57	94.74
DSP6	Brandenburg	47.69	8.57	128	10.34	136	40.44
DSP3	Brandenburg	27.52	1.80	146.5	1.91	80	93.75
DSP8	Brandenburg	65.18	5.65	122	5.18	31	58.06
DSP2	Brandenburg	91	4.63	64	4.97	71	60.56

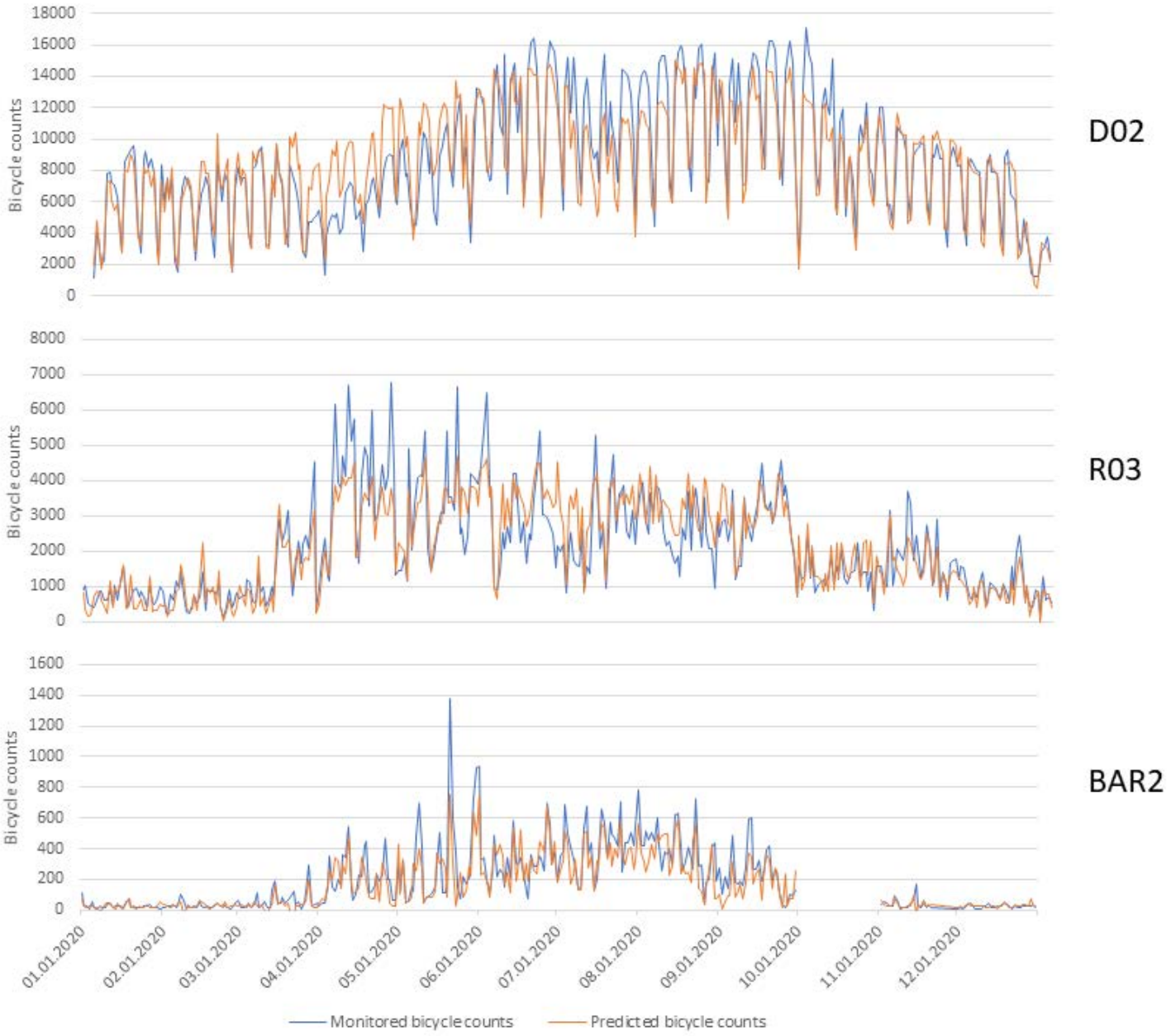

Figure 5: Predicted vs. monitored bicycle counts of the unknown year model.

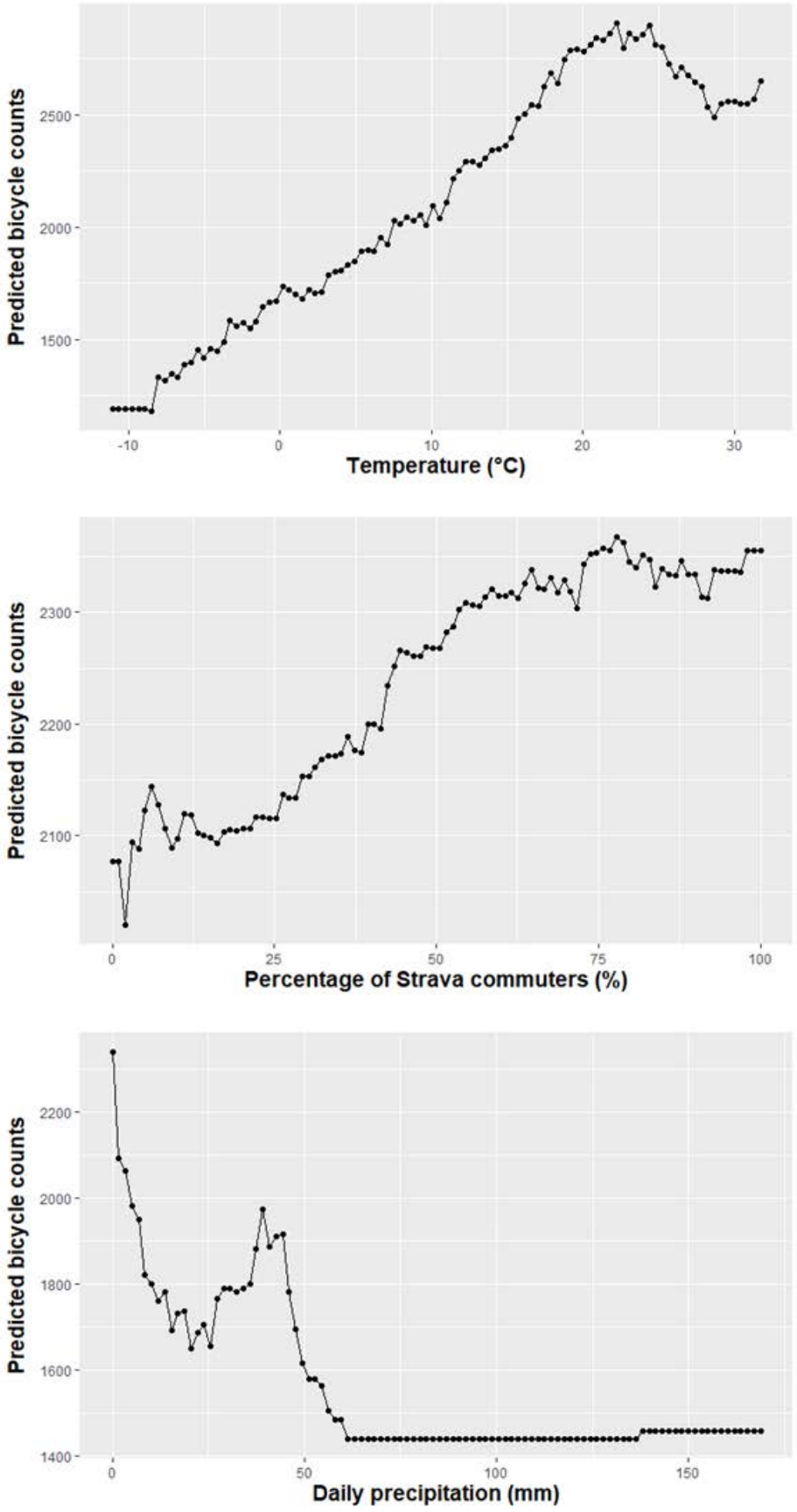

Figure 6: GBM prediction of bicycle counts for increasing temperature, percentage of Strava commuters and precipitation.

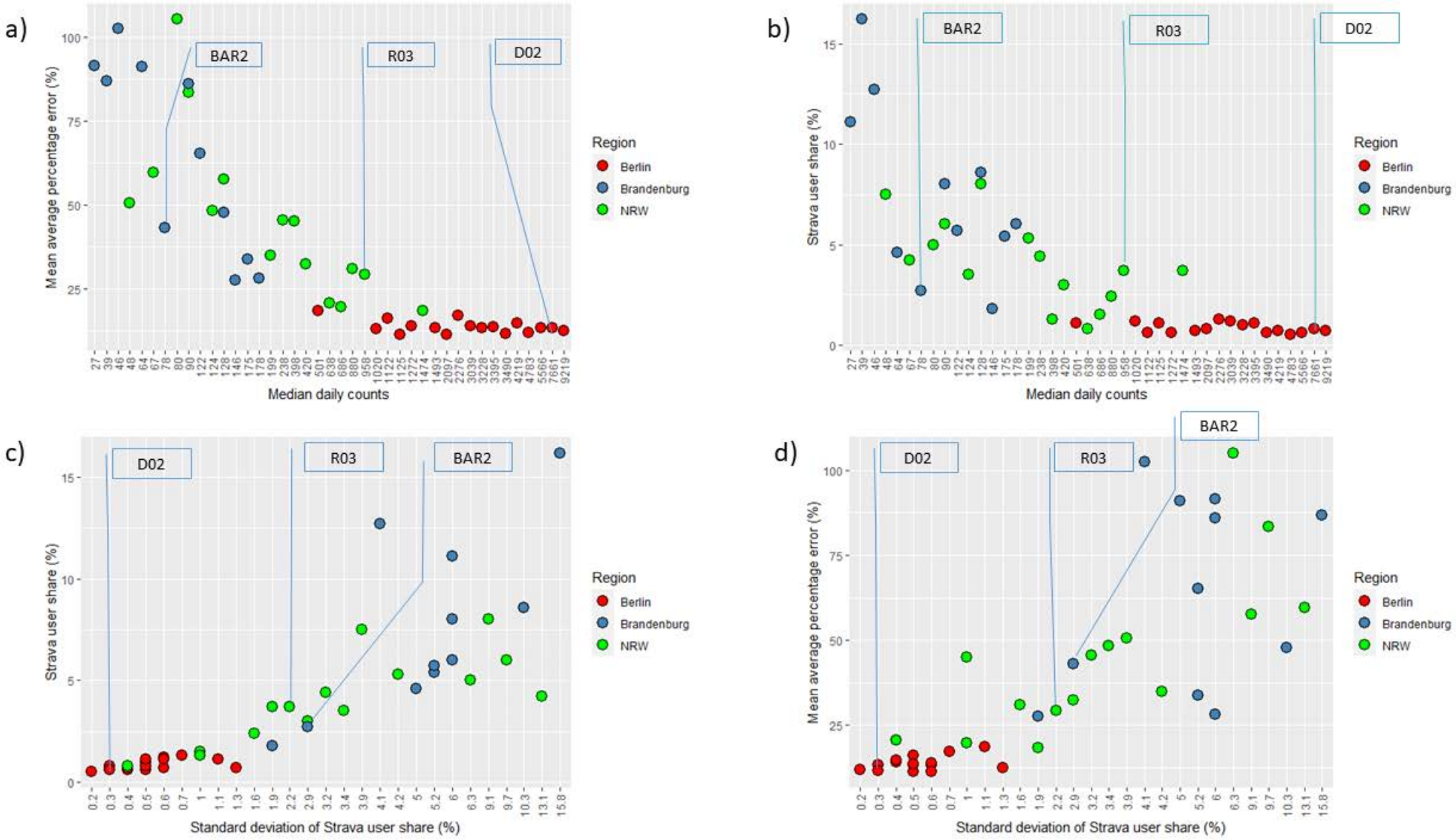

Figure 7:

a) *Median daily counts sorted in ascending order versus MAPE of the standard model estimations for each counter location.*

b) *Median daily counts sorted in ascending order versus Strava user share for each counter location.*

c) *Standard deviation of Strava user share in ascending order versus MAPE of the standard model estimations for each counter location.*

d) *Standard deviation of Strava user share in ascending order versus Strava user share for each counter location.*

Table 8: Mean Strava user share and standard deviation for monitored count quantiles.

Monitored count range	Mean Strava user share (%)	Standard deviation of Strava user share
≤ 87	15.10	15.44
87 - 234	5.27	3.46
234 - 496	3.20	2.18
496 - 917	1.90	1.69
917 - 1608	1.30	1.12
1608 - 2683	1.11	0.90
28683 - 4837.125	1.00	0.77
≥ 4837.125	0.69	0.50

Table 9: Mean Strava user share and standard deviation for monitored Strava count quantiles.

Strava count range	Mean Strava user share (%)	Standard deviation of Strava user share
≤ 10	2.10	4.62
10 - 20	1.99	3.47
20 - 35	1.63	2.43
≥ 35	1.73	1.65

9. References

Al-Ramini, A., Takallou, M. A., Piatkowski, D. P., & Alsaleem, F. (2022). Quan tifying changes in bicycle volumes using crowdsourced dat a. *Environment and Planning B: Urban Analytics and City Science*, 23998083211066103.

Berlin-Brandenburg, A. f. S. (2021). *Einwohnerinnen und Einwohner in den Ortsteilen Berlins am 31.12.2020*. https://daten.berlin.de/datensaetze/einwohnerinnen-und-einwohner-den-ortsteilen-berlins-am-31122020

BKG, B. f. K. u. G. (2018). *Landbedeckungsmodell für Deutschland (LBM-DE) Geobasisdaten: © GeoBasis-DE / BKG* [Land Use]. https://gdz.bkg.bund.de/index.php/default/digitales-landbedeckungsmodell-fur-deutschland-stand-2018-lbm-de2018.html

Bossard, M., Feranec, J., & Otahel, J. (2000). *CORINE land cover technical guide: Addendum 2000* (Vol. 40). European Environment Agency Copenhagen.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData mining*, *10*(1), 1-17.

ClimateData. (2022a). *Berlin* Climate Data. https://en.climate-data.org/europe/germany/berlin/berlin-2138/

ClimateData. (2022b). *Dortmund* Climate Data. https://en.climate-data.org/europe/germany/north-rhine-westphalia/dortmund-147/

Conrow, L., Wentz, E., Nelson, T., & Pettit, C. (2018). Comparing spatial patterns of crowdsourced and conventional bicycling datasets. *Applied geography*, *92*, 21-30.

Cornes, R., G. van der Schrier, E.J.M. van den Besselaar, and P.D. Jones. (2018). *An Ensemble Version of the E-OBS Temperature and Precipitation Datasets*. https://doi.org/10.1029/2017JD028200

Dadashova, B., & Griffin, G. P. (2020). Random parameter models for estimating statewide daily bicycle counts using crowdsourced data. *Transportation Research Part D: Transport and Environment*, *84*. https://doi.org/10.1016/j.trd.2020.102368

El Esawey, M., Mosa, A. I., & Nasr, K. (2015). Estimation of daily bicycle traffic volumes using sparse data. *Computers, Environment and Urban Systems*, *54*, 195-203.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *J Anim Ecol*, *77*(4), 802-813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Eurostat-Data. (2022). *List of Urban-rural regions (NUTS-2021)* [Table]. https://ec.europa.eu/eurostat/documents/345175/629341/NUTS2021.xlsx

Eurostat. (2022). Urban-Rural Typology. Retrieved 09.11.2022, from https://ec.europa.eu/eurostat/web/rural-development/methodology

Flaig, B. B., & Barth, B. (2014). Die Sinus-Milieus® 3.0–Hintergründe und Fakten zum aktuellen Sinus-Milieu-Modell. In *Zielgruppen im Konsumentenmarketing* (pp. 105-120). Springer.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Griffin, G. P., & Jiao, J. (2019). Crowdsourcing Bicycle Volumes: Exploring the role of volunteered geographic information and established monitoring methods.

Hankey, S., Lindsey, G., Wang, X., Borah, J., Hoff, K., Utecht, B., & Xu, Z. (2012). Estimating use of non-motorized infrastructure: Models of bicycle and pedestrian traffic in Minneapolis, MN. *Landscape and Urban Planning*, *107*(3), 307-316.

Hankey, S., Lu, T., Mondschein, A., & Buehler, R. (2017). Spatial models of active travel in small communities: Merging the goals of traffic monitoring and direct-demand modeling. *Journal of Transport & Health*, *7*, 149-159.

Hering, M.-M. (2019). Baustelle an der Oberbaumbrücke: Fußgänger-Lobby kritisiert Umleitung auf "Ekelweg" [Online Article]. Retrieved 03.08.2022, from https://www.tagesspiegel.de/berlin/fussganger-lobby-kritisiert-umleitung-auf-ekelweg-6584536.html

Jestico, B., Nelson, T., & Winters, M. (2016). Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography*, *52*, 90-97. https://doi.org/10.1016/j.jtrangeo.2016.03.006

Kamalapuram, S. (2022). *Estimating bicycle and pedestrian ridership using the Random Forest algorithm* UNIVERSITY OF CALIFORNIA DAVIS].

Lee, K., & Sener, I. N. (2021). Strava Metro data for bicycle monitoring: a literature review. *Transport Reviews*, *41*(1), 27-47.

Livingston, M., McArthur, D., Hong, J., & English, K. (2021). Predicting cycling volumes using crowdsourced activity data. *Environment and Planning B: Urban Analytics and City Science*, *48*(5), 1228-1244.

Miah, M., Mattingly, S., Hyun, K., & Broach, J. (2021). Challenges and Opportunities of Emerging Data Sources to Estimate Networkwide Bike Counts.

Miah, M. M., Hyun, K. K., Mattingly, S. P., & Khan, H. (2022). Estimation of daily bicycle traffic using machine and deep learning techniques. *Transportation*, 1-54.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885-900.

Nelson, T., Roy, A., Ferster, C., Fischer, J., Brum-Bastos, V., Laberee, K., Yu, H., & Winters, M. (2021). Generalized model for mapping bicycle ridership with crowdsourced data. *Transportation Research Part C: Emerging Technologies*, *125*, 102981.

Nordrhein-Westfalen, I. u. T. (2022). *Bevölkerungsstand nach 5er- Altersgruppen (19) und Geschlecht - Gemeinden - Stichtag*. https://www.landesdatenbank.nrw.de/link/tabelleDownload/12411-06iz

OpenStreetMap. (2022). *Planet dump retrieved from https://planet.osm.org*. https://www.openstreetmap.org

Pérez-Sánchez, M., Sánchez-Romero, F.-J., Ramos, H. M., & López Jiménez, P. A. (2017). Calibrating a flow model in an irrigation network: Case study in Alicante, Spain. *Spanish Journal of Agricultural Research (Online)*, *15*(1), 1-13.

Pucher, J., & Buehler, R. (2017). Cycling towards a more sustainable transport future. *Transport Reviews*, *37*(6), 689-694.

Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package. *Update*, *1*(1), 2007.

Roy, A., Nelson, T. A., Fotheringham, A. S., & Winters, M. (2019). Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists. *Urban Science*, *3*(2). https://doi.org/10.3390/urbansci3020062

Statistisches_Bundesamt. (2011). *Einwohnerzahl je Hektar*. https://www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html

Strava_Inc. (2020).

Venohr, M., Langhans, S. D., Peters, O., Hölker, F., Arlinghaus, R., Mitchell, L., & Wolter, C. (2018). The underestimated dynamics and impacts of water-based recreational activities on freshwater ecosystems. *Environmental Reviews*, *26*(2), 199-213.

Watkins, K., Ammanamanchi, R., LaMondia, J., & Le Dantec, C. A. (2016). *Comparison of Smartphone-based Cyclist GPS Data Sources*