

APPLICATION PAPER 

From counting stations to city-wide estimates: data-driven bicycle volume extrapolation

Silke K. Kaiser^{1,2} , Nadja Klein³  and Lynn H. Kaack^{1,2} 

¹Data Science Lab, Hertie School, Berlin, Germany

²Center for Sustainability, Hertie School, Berlin, Germany

³Scientific Computing Center, Karlsruhe Institute of Technology, Karlsruhe, Germany

Corresponding author: Silke K. Kaiser; Email: s.kaiser@phd.hertie-school.org

Received: 13 March 2024; **Revised:** 16 December 2024; **Accepted:** 15 January 2025


Keywords: bicycle volume; climate policy; machine learning; sustainable transportation

Abstract

Shifting to cycling in urban areas reduces greenhouse gas emissions and improves public health. Access to street-level data on bicycle traffic would assist cities in planning targeted infrastructure improvements to encourage cycling and provide civil society with evidence to advocate for cyclists' needs. Yet, the data currently available to cities and citizens often only comes from sparsely located counting stations. This paper extrapolates bicycle volume beyond these few locations to estimate street-level bicycle counts for the entire city of Berlin. We predict daily and average annual daily street-level bicycle volumes using machine-learning techniques and various data sources. These include app-based crowdsourced data, infrastructure, bike-sharing, motorized traffic, socioeconomic indicators, weather, holiday data, and centrality measures. Our analysis reveals that crowdsourced cycling flow data from Strava in the area around the point of interest are most important for the prediction. To provide guidance for future data collection, we analyze how including short-term counts at predicted locations enhances model performance. By incorporating just 10 days of sample counts for each predicted location, we are able to almost halve the error and greatly reduce the variability in performance among predicted locations.

Impact Statement

We show how data science can be used to achieve urban sustainability goals at the nexus of climate change and health by promoting active transportation. Our work demonstrates how bicycle volume can be extrapolated from a few scarcely located counting stations to street-level predictions. Such spatial extrapolation based on urban traffic sensor data has received little scholarly attention, especially in the case of bicycles. We generate predictions with machine learning approaches using a wide range of different data sources. Among others, we employ bike-sharing data that we scraped and made available and illustrate how the different data modalities can be feature-engineered. Providing this granularity of cycling data is a necessary step towards evidence-based infrastructure development, a key factor in promoting cycling.

 This research article was awarded Open Data badge for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

1. Introduction

Shifting from motorized transport to bicycles offers significant health and environmental benefits, including improved cardiorespiratory health, reduced cancer mortality risk (Oja et al., 2011; Woodcock et al., 2009), and lower greenhouse gas emissions (Pörtner et al., 2022). Enhancing bicycle infrastructure is a promising strategy to encourage cycling in urban areas. Research indicates that cyclists, particularly women, prefer dedicated bike infrastructure (Dill, 2009; Garrard, Rose, and Lo, 2008). Also, riding in a separate bicycle lane is linked to a reduced risk of accidents (Morrison et al., 2019). Introducing new bike lanes, however, is often highly contested due to limited resources, such as funding and road space. Thus, data-driven approaches are crucial for accurately targeting infrastructure improvements in areas with the greatest need (Olmos et al., 2020; Larsen, Patterson, and El-Geneidy, 2013).

One relevant piece of information for such data-driven approaches is bicycle volume data. Currently, most of this data is collected by permanently installed bicycle counting stations, providing information on cyclists passing by a specific location. Due to their high cost, these stations are sparsely located across a road network. At the same time, several data sources related to cycling are openly available (Romanillos et al., 2016). Given the scarcity of bicycle volume data on the one hand and the abundance of related data on the other hand, clamors for methods that are able to make use of all available information in order to better predict bicycle volumes at a fine-grained scale. We address this by combining machine learning (ML) methods with a wide variety of available data sources to extrapolate bicycle volume to a much higher spatial resolution. With this machinery, we aim to answer three important questions. First, can we predict bicycle volume at unseen locations using a variety of data? Second, which of these data sources are the most relevant for prediction? And third, how much can the performance be improved by adding sample counts for the predicted locations?

Researchers have identified several datasets related to bicycle volume that have proven useful, especially for interpolating missing observations in bicycle count data. These include data sources that have long been available, such as weather, holidays, infrastructure, and socioeconomic indicators (Miranda-Moreno and Nosal, 2011; Strauss and Miranda-Moreno, 2013; Holmgren, Aspegren, and Dahlströma, 2017). More recently, the growing availability of data associated with widespread smartphone use has opened new avenues for analysis (Lee and Sener, 2020). Notably, this includes valuable information from crowdsourced bicycle usage data, in particular, from the Strava application (Lee and Sener, 2021; Kwigizile, Morgan Kwayu, and Oh, 2022), bike-sharing protocols (Miah et al., 2023) or the use of photos and tweets (Wu et al., 2017).

Among available studies, some extrapolate bicycle volume using only a few of these data sources. For instance, Miah et al., 2022 explore how counting station data can be merged with crowdsourced data to estimate bicycle volumes across street networks using clustering and nonparametric modeling. They find that relying solely on crowdsourced data as an additional input to counts is challenging, particularly due to oversampling from counting stations located at high-volume locations. Similar studies estimate cyclists' exposure employing various data sources and using classical regression approaches (Sanders et al., 2017; Griswold, Medury, and Schneider, 2011), mixed effects models (Dadashova and Griffin, 2020) or Poisson regressions (Roy et al., 2019). In addition to traditional statistical approaches, ML methods have been increasingly applied over the past decade. For instance, (Sekula et al., 2018; Das and Tsapakis, 2020; Zahedian et al., 2020) have proven how ML methods can be leveraged for the extrapolation of motorized traffic. However, to the best of our knowledge, there is no study that combines ML methods with a large variety of different data sources to provide reliable, fine-grained predictions of bicycle counts beyond available counting stations.

Our paper showcases our approach in the city of Berlin. In Germany's largest city, with 3.6 million inhabitants, the modal share for walking and cycling was 37% in 2023, which is above the European average of 33% (based on 31 major cities with a minimum of 13% and a maximum of 57%) (European Metropolitan Transport Authorities, 2023). We note that each city has unique characteristics, and while Berlin's case promises to provide valuable insights, it cannot represent the diversity of urban settings across Europe, thereby limiting the generalizability of our findings to other cities. We implement and compare different ML algorithms to predict the daily and average annual daily bicycle volume (AADB) at unseen locations. We use a wide array of features, many of which have proven pertinent in previous

Table 1. Overview of data types used in this paper to predict bicycle volume and their use in other publications: including crowdsourced (Strava)[Crowds.], infrastructure [Infr.], weather [Weath.], socioeconomic [Socio.], bike-sharing [B.-S.], public and school holidays [Hol.], centrality measures [Centr.], and motorized traffic [Moto.]

Reference/data source	Crowds.	Infr.	Weath.	Socio.	B.-S.	Hol.	Centr.	Moto.
Miah et al. (2023)	✓	✓	✓	✓	✓			
Dadashova and Griffin (2020)	✓	✓	✓	✓				
Kwigizile et al. (2022)	✓	✓	✓	✓				
Hochmair et al. (2019)	✓	✓		✓			✓	
Sanders et al. (2017)	✓	✓		✓				
Nelson et al. (2021)	✓		✓	✓				
Hankey and Lindsey (2016)		✓	✓	✓				
Strauss and Miranda-Moreno (2013)		✓	✓	✓				
Roy et al. (2019)	✓	✓						
El Esawey (2018)			✓					
Holmgren et al. (2017)			✓			✓		
Miranda-Moreno and Nosal (2011)			✓					
Lu et al. (2017)		✓					✓	
This paper	✓	✓	✓	✓	✓	✓	✓	✓

studies (see Table 1 for an overview). To identify the most relevant data sources, we perform a grouped base permutation feature importance. Lastly, we aim to guide future data collection by evaluating whether collecting sample counts at unseen locations would be purposeful to improve the predictions further, and what is the best strategy to collect this data.

2. Results

2.1. Data sources

Our study uses data from 20 long-term bicycle counting stations in Berlin, which continuously measure the number of passing bicycles per hour. In addition, we employ data from 12 short-term counting stations, where counts are conducted on individual days throughout the year (Senate Department for the Environment, Mobility, Consumer and Climate Protection Berlin, 2022). To accurately predict bicycle counts, we make use of information contained in a variety of further sources. These include data on infrastructure, socioeconomic factors, motorized traffic, weather, holidays, centrality measures, bike-sharing, and from a crowdsourcing application that tracks cyclists (Strava application). We also use inherent information on the time. Bike-sharing and Strava data directly represent bicycle traffic. However, they attract different users and differ in the type of information they provide. The former describes the exact time and origin–destination-pairs of individual trips taken on short-term free-floating rented bikes. The latter are anonymized georeferenced data from an application, which are aggregated to provide the number of trips for a region and for road segments between intersections based on tracking users as they ride. The bike-sharing, crowdsourced, and motorized traffic data are feature-engineered, to indicate the usage volume, respectively of passing motorized traffic within different radii around counting stations. The socioeconomic and infrastructure features are assigned in accordance with the location of the counting stations. Further details on the distinct data sources, including data clearing and feature engineering are provided in the Methods Section 4.1. A list of all features is provided in Table 2 together with references to the data sources. The bike-sharing data is only available for April to December 2019 and June to December 2022. Therefore, we set our study period to these periods. This also largely omits the period of the COVID-19 pandemic and its impact on transportation.

Table 2. Overview of the features per data source used in this study

Data category	Description of features	# of features	Data source
Crowdsourced	Number of trips originating, arriving, or happening; with respect to leisure and commute, with respect to different times of the day, with respect to the weekend, with respect to different personal characteristics (age, sex), with respect to normal and e-bikes, as well as average speed (both for hexagon and street segment data)	135	Strava Metro (2023)
Infrastructure	Latitude, longitude, distance to city center, maximum speed, bicycle lane type, number of shops/education centers/hotels/hospitals/industries for various radii, percent of area used for farming/horticulture/cemeteries/waterways/industry/ private gardening/parks/traffic areas/forests/ residential housing	50	OpenStreetMap contributors (2017) and Senate Department for Urban Development, Building and Housing, (2023)
Weather	Average/maximum/minimum temperature, precipitation, maximum snow depth, sunshine duration, wind speed, wind direction, peak wind gust, dew point, air pressure, humidity	10	meteostat (2022)
Socioeconomic	Population density, total number of inhabitants, average age, gender distribution, share of population with migration background, share of foreigners, share of unemployed, share of population with tenure exceeding 5 years, rate moving to/from area, age-specific demographic proportions, greying index, birth rate	15	Senate Department for Urban Development, Building and Housing (2023) and Berlin-Brandenburg Office of Statistics (2023)
Bike-sharing	Number of bicycles originated, returned rented within various radii	24	CityLab Berlin (2019), Nextbike (2020) and Kaiser (2023)
Holiday	School holiday, public holiday	2	Senate Department for Education, Youth and Family (2022)
Motorized traffic	Total number and speed of vehicles/cars/lorries within different radii	12	Berlin Open Data (2022b)
Centrality measures	Degree, closeness, betweenness, clustering coefficient	4	based on Berlin Open Data (2024)
Time	Month, day of month, weekday, weekend, year	5	Inherent
Total number of features		257	

2.2. Spatial extrapolation using multi-source data

We train our model using data from existing counting stations as ground truth. We compare the performance of different ML algorithms on this task. A description of the models, the feature selection, and the hyperparameter tuning can be found in the Methods Section 4.2. We evaluate the predictions on the daily and average annual (AADB) scales. The daily scale is valuable for providing a more detailed picture of the variation throughout the year, and it is relevant for understanding the effects of intra-week variation, special events, and seasonal weather conditions (Yi et al., 2021; Sekuła et al., 2018; Zahedian et al., 2020). For infrastructure planning decisions, annual averages may be sufficient. The AADB is the average number of bicycles that pass a given location per day for a given year. We compute the performance for the AADB by predicting the daily counts and evaluating their average against the annual ground truth average. Since the counting station data is recorded hourly, we sum up the measurements for each day to obtain daily measurements. To address non-normal distribution characterized by a pronounced right skew of the ground truth, we apply a logarithmic transformation (see Methods Section 4.2). To simulate extrapolation, we evaluate our models using leave-one-group-out (LOGO) cross-validation (CV). The method follows the same principle as standard CV but differs in how the data is partitioned. Instead of random partitioning, the data is organized into distinct groups, which, in our case, correspond to counting stations. Consequently, the model is trained on observations from all but one long-term counting station and then evaluated on this hold-out long-term counting station.

In addition, we use the short-term counting locations as test data for a model trained on all long-term counting stations. We provide the average error across stations, which implies that each location is equally weighted in the test data. When computing these predictions, it is important to note that the hourly long-term data are measured from 0 h to 24 h, while the short-term counts only from 7 h to 19 h. Hence, we train the model, predicting the short-term locations, only on daily measurements, which are computed as the sum of the 7 h–19 h hourly measurements. We also perform the analysis of the long-term stations on daily measurements based on 0 h–24 h and 07 h–19 h data separately. The former allows us to infer day effects for long-term stations, and the latter can be used to compare results with the short-term counting predictions.

In order to provide information on the absolute and relative size of our errors, we use the mean absolute error (MAE), and the symmetric mean absolute percentage error (SMAPE) as evaluation metrics and train the models on various ML algorithms (see Methods Section 4.2). Additionally, we include a baseline for comparison, where predictions are generated using the mean of the observations in the training data.

We find that all models, besides the linear regression, outperform the baseline in all specifications. XGBoost also outperforms decision trees, random forests, support vector machines, linear regression, and shallow neural networks in all specifications (Table 3). For reasons of brevity, we chose one model to conduct the subsequent analysis. As the analysis deals with long-term counting stations on the 0 h–24 h data, we select XGBoost, due to its slightly superior performance in terms of both MAE and SMAPE for this specification. In particular, we note that this model does not produce the smallest MAE in predicting the long-term counts for the 7–19 h data. The choice of XGBoost is made in the context of this specific use case and is not a general recommendation for this model. To analyze the performance of the XGBoost model in more detail, we looked into the variation of SMAPE between stations. At the daily scale, the model performs quite well for more than half the stations (SMAPE of up to 30), while for some, the SMAPE exceeds 80 (Figure 1a), and the performance also varies considerably between counters for the AADB (Figure 1b). The poorly performing locations each have a high variance in their measurements, and each of these locations is either consistently over-predicted or under-predicted. Our analysis revealed no further common characteristics of the worst-performing counters that would allow us to pinpoint where the model is failing. We conclude that there are latent factors within the data generation process that remain unaccounted for despite our comprehensive inclusion of a wide range of features from the existing literature. We will explore how this can be mitigated using sample counts in Section 2.5.

2.3. Relative importance of feature groups

Each data source used requires time and effort for acquisition, cleaning, and integration. Given the variety of sources used in this study, we explore their relative importance so that individuals considering a similar modeling approach can anticipate which ones are essential to obtain.

Table 3. Errors for the various machine learning models at the daily, and average annual daily bicycle volume (AADB) scale. The gray background implicates the columns employed as the criterion for model selection for the subsequent analysis

(a) MAE						
Dimension	Daily	Daily	Daily	AADB	AADB	AADB
Time	0 h–24 h	7 h–19 h	7 h–19 h	all day	7 h–19 h	7 h–19 h
Counter type	Long-term	Long-term	Short-term	Long-term	Long-term	Short-term
Evaluation	LOGO	LOGO	Test	LOGO	LOGO	Test
	(1)	(2)	(3)	(4)	(5)	(6)
Linear regression	2275.83	1681.07	1803.38	2090.51	1500.85	1709.86
Decision tree	2139.44	1591.54	1425.18	1804.12	1327.36	1464.00
Random forest	1630.23	1609.00	1050.08	1477.03	1438.34	1067.96
Gradient boosting	1760.96	1477.24	1149.74	1575.33	1295.48	1132.62
XGBoost	1511.58	1540.70	828.94	1342.96	1339.80	913.75
Support vector machine	1793.84	1612.20	1225.12	1668.06	1404.77	1019.57
Shallow neural network	2186.43	1907.65	1611.37	1637.59	1816.15	1688.72
Baseline	2397.87	1944.66	1707.61	2397.87	1944.66	1707.61

(b) SMAPE						
Dimension	Daily	Daily	Daily	AADB	AADB	AADB
Time	0 h–24 h	7 h–19 h	7 h–19 h	0 h–24 h	7 h–19 h	7 h–19 h
Counter type	Long-term	Long-term	Short-term	Long-term	Long-term	Short-term
Evaluation	LOGO	LOGO	Test	LOGO	LOGO	Test
	(1)	(2)	(3)	(4)	(5)	(6)
Linear regression	62.91	51.78	68.39	61.26	49.17	65.51
Decision tree	47.24	46.00	60.27	40.19	38.81	58.78
Random forest	38.18	48.72	47.57	35.40	46.76	47.02
Gradient boosting	45.20	45.05	54.82	42.83	42.41	52.35
XGBoost	37.36	44.52	43.24	34.26	42.47	37.89
Support vector machine	47.68	52.79	56.93	48.36	47.94	50.18
Shallow neural network	47.77	57.46	67.58	55.88	53.38	65.75
Baseline	64.10	63.14	69.80	64.10	63.14	69.80

Feature importance measures the contribution of the feature to the prediction of the target variable. Given the large number of employed features and possible collinearity among them, we evaluate their grouped importance. We group them along the data sources, further splitting them to account for their different characteristics, such as whether they provide information on flows, travel demand, origin–destination information, and the different radii used (particularly local vs. whole city features). With this approach, collinear features are included in the same group, ensuring that their combined contribution is evaluated rather than wrongly estimating the importance of any single feature. We only use those features that the feature selection method selected for the XGBoost model. Our categories include time, holidays, and weather derived from the data sources. We further divide motorized traffic features into city-wide and

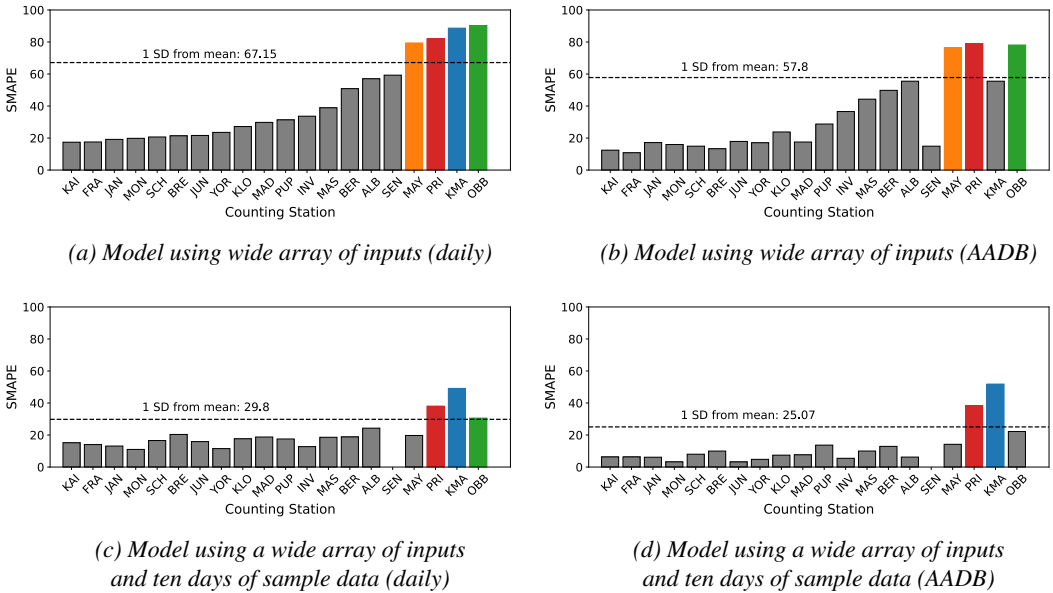
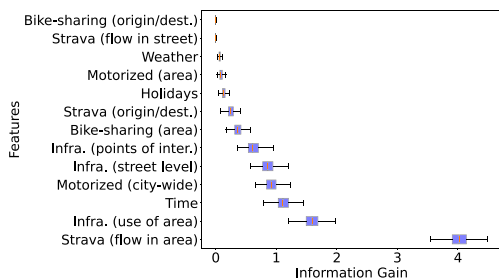


Figure 1. Performance of XGBoost model at the daily level and for average annual daily bicycle volume estimations (AADB) across the individual counting stations. Subfigure b) and d) were trained on 10 days’ worth of sample data and on the additional long-term counting stations (full-city model specification). Highlighted in all graphs are the counting stations whose error exceeds or is below a deviation of 1 standard deviation from the mean. The color coding and the ordering of the counting stations across all subplots are the same to ensure comparability. The counting station ‘SEN’ is left out in subplot b) and d), due to the small number of observations available.

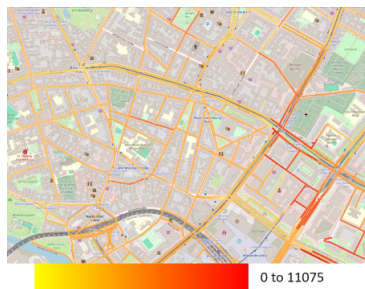
area-specific groups. Infrastructure features include street-level features (speed, type of bicycle lane), points of interest (such as the number of shops or educational centers), and features related to the usage of the area around the counting location. Bike-sharing features are grouped into those capturing the count of passing bicycles and those tracking the number of trips that start or end. Lastly, we divided Strava data into flow at different spatial levels (area-wide, city-wide, and street-level), as well as origin/destination features. The feature selection process did not include any socioeconomic or centrality features. A detailed list of the selected features and their assignment to the groups is included in [Appendix A.1](#).

We compute the grouped feature importance at the daily scale, using the Grouped Permutation Importance (GPI; Plagwitz et al., 2022), which is described in [Section 4.2](#). Additionally, we focus on the SMAPE error, as correctly predicting both relatively busy and relatively slow roads is valuable when deciding where to prioritize infrastructure. Finally, since we want to get a comprehensive picture of the daily traffic situation, including at night, we use the data for the 0 h–24 h time window. We train the model on all long-term counting stations. Within GPI, we compute 100 permutations and use repeated 5-fold stratified CV.

The GPI reveals that crowdsourced Strava application data, specifically those features describing the flow in the area around the point of interest, is the most important group, followed by infrastructure (use of area), time indicators, motorized traffic (city-wide), infrastructure (street-level) and infrastructure (points of interest) ([Figure 2a](#)). The crowdsourced information is much more relevant than the bike-sharing data. While both directly represent bicycle traffic, the movement patterns of individuals tracking their trips turn out to be more indicative of the overall cycling volume. Therefore, consistent with previous research, we find that Strava indicators are very useful for estimating cycling volumes (Sanders et al., 2017; Hochmair, Bardin, and Ahmouda, 2019; Kwigizile, Morgan Kwayu, and Oh, 2022).



(a) Information gain computed via a grouped permutation importance for the XGBoost model at the daily level using SMAPE.



(b) Proof of concept: Application of the XGBoost model to a subset of Berlin streets to predict the streetwise daily bicycle volume. The prediction in the picture is for 20.09.2022.

Figure 2. Feature importance and proof of concept based on an XGBoost model trained on data of all available long-term counting stations.

2.4. Proof of concept of multi-source model

We empirically demonstrate benefits from our multi-source model by simulating daily streetwise bicycle volume in a subarea of Berlin for the month of September 2022. Figure 2b shows a snapshot from the simulation, which is available online at <https://silkekaiser.github.io/research>. To generate these predictions, we used XGBoost, trained on data from all available long-term counting stations. Specifically, we predict the bicycle volume for each street segment between two intersections, using the midpoint of each segment as the reference point for our estimates.

We find that the demonstration effectively captures temporal variations, especially between weekends and weekdays. However, the spatial aspects of the predictions could be more convincing. The model often predicts that adjacent streets have similar bicycle volumes and fails to detect high values. This former shortcoming is likely due to the construction of features based on large radii. The usage of the log transformation of the target variable might amplify the latter. Nevertheless, the model reasonably captures the differences between major streets and residential areas, picking out high and low-traffic zones.

2.5. Spatial extrapolation using additional sample count data

Our multi-source model has only a limited ability to reproduce spatial patterns of cycling volume. Here, we investigate whether collecting additional location-specific bicycle volume sample counts improves the predictive performance at unseen locations on a daily scale and what is the most effective strategy for conducting them.

Hankey, Lindsey, and Marshall, 2014 elaborate on the usefulness of short-term counts to estimate annual averages for non-motorized traffic using scaling factors. They find that as the number of observation days increases, the extrapolation error decreases, but that the incremental gains become modest after the first week. Also, the advantage derived from conducting counts on consecutive days is minimal compared to nonconsecutive days. We seek to revisit their findings in the context of ML. We chose to simulate three different sample data collection strategies: Firstly, the collection of data is commissioned for 1 day at a time (1-day). The days are selected at random throughout the year. In the second and third strategies, we simulate the collection of data on three (3-day) or seven (7-day) consecutive days. Also, these multi-day periods are randomly distributed throughout the year. We compare the performance of the model with data from each of those three different sampling strategies. We simulate a collection of up to 28 days.

We simulate this using 19 long-term counting stations only, as all short and one long-term station have too few observations available. We employ the XGBoost model with SMAPE, using only the features selected by the feature selection, just as in Section 2.2. As before, we evaluate the performance by iterating

over the counting stations. Each counting station serves once as the new (“hold-out”) location. For that location, we randomly choose some of the available observations to represent sample counts performed at that location, following the three sampling strategies (1, 3, or 7-day). We use the remaining data from the location as the test set. For training, we implement two scenarios. For the first scenario, we make use of all available data: we train the model on both the sampled observations and all the observations from the other counting stations. We assign weights to the data, giving 25% weight to the sample counts and 75% to the observations from other counting stations. Please refer to the Methods Section 4.2 for details on the weights. This “full-city” scenario benefits from both location-specific sample data and city-wide long-term information. For the second scenario, we train the model only on the sample data. Since it only uses information from the location in question, we refer to this model as the “location-specific” scenario. Thus, by definition, the training data for this model exhibits no variation in infrastructure and socioeconomic features, as these features only vary across locations. We then use both scenario models to perform predictions on the test set. We repeat this process for each counting station and compute the average across the resulting errors. This procedure is repeated 10 times with different sample days to allow for uncertainty estimation. We train and evaluate the models after each additional day of data collection. This allows a comparison of the different approaches for as little as 1 day and as much as 28 days of additionally sampled data (see Appendix A.2 for a graphical representation of this full methodology). As a simple baseline, we include the error of predicting the site-specific volume as the mean of the sample data collected at the respective location.

Sample data collection notably enhances predictive performance for new locations in the full-city scenario (Figure 3). In the location-specific scenario, two or more days’ worth of sample data already outperforms a model without any location-specific data. Sampling only 1 day at a time is the superior collection strategy for both scenarios, and this advantage is more pronounced for the full-city scenario. Collecting data on as many different days as possible may provide an advantage, as seasonal effects are better captured. Given that setting up counting infrastructure at new locations may be costly, the 3-day and 7-day approaches may still yield sufficient results at lower costs. Moreover, we find that the full-city approach using the 1-day strategy outperforms the location-specific approaches up until the 8th day of data collection (across all repeated samples) and in the mean also thereafter. A comparison of the 1-day

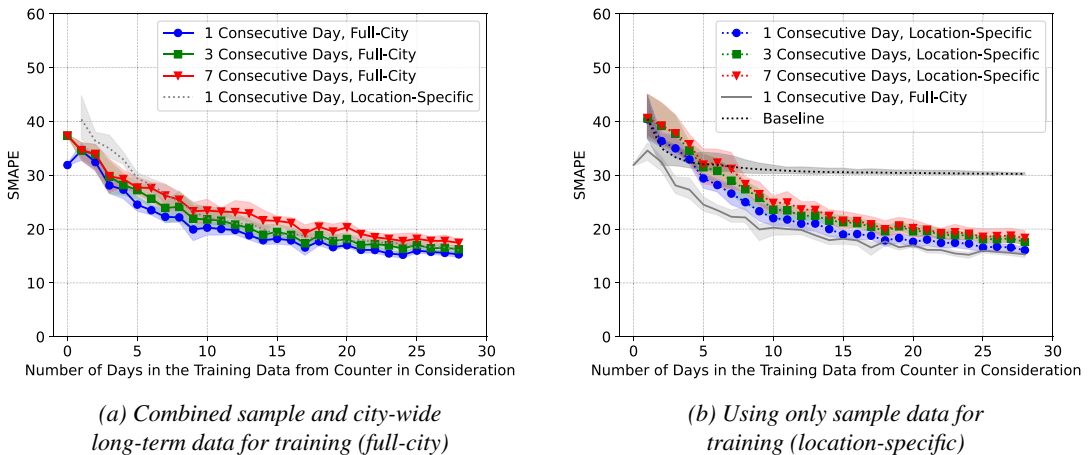


Figure 3. The effect of collecting additional sample data at a new location to predict the daily volume of bicycles using XGBoost. In the left diagram, the models are trained on the full-city available data, both long-term data from other sites and sample data from the location in question; in the right diagram, the models are trained on location-specific sample data only. Best-performing specifications are depicted in gray in the other plot to allow for comparison. The error is the average over the 19 counting stations used, with 95% confidence intervals calculated from 10 repeated samples.

strategy between the two approaches shows that to achieve a SMAPE of 20, one would need to collect on average 9 days of sample data using the full-city scenario or 14 days using the location-specific model. This underscores the fact that models can benefit greatly from information obtained at locations other than the one under consideration. Finally, we find that the use of multi-source data is also highly relevant when using sample data, and simple averages over the counts do not suffice. The baseline error never drops below 30, while the errors for the location-specific models are below 20 after 15 days of sample counts (Figure 3b). This demonstrates the importance of leveraging multi-source data in combination with sample counts.

Based on these results, we seek to provide a numerical comparison of the performance of a model with sample data to the simple multi-source model. We compute the full-city model using 10 days' worth of sample counts using the 1-day strategy. With this approach, we can predict new locations at the daily scale with an average SMAPE of 20.22 (in comparison to the multi-source model of 37.36) and an MAE of 641.57 (1511.58). For the AADB, we get 11.85 (34.26) and 416.81 (1342.96), respectively. On closer inspection, we also find that these errors vary little across counting stations (Figure 1b and 1d). This is a clear improvement over the multi-data-only model. Therefore, estimates predicted with sample counts and multi-source data are not only more accurate (depending on the specification, almost halved to 2/3 lower) but also more reliable.

3. Discussion

Our research highlights the feasibility of estimating bicycle volumes for all streets across a city by leveraging open-source data together with long- and short-term counting station data and machine-learning models. Advances not only in data availability but also in analytical methods have made such purely data-driven approaches feasible.

We find that leveraging already existing multi-source and long-term counting station data allows for predicting bicycle volume at unseen locations using XGBoost with a reasonable error for both daily values and annual averages (SMAPE of 37.36 and 34.26, respectively). Direct comparison with previous studies is limited due to differences in focus and error metrics. Studies that focused mainly on temporal interpolation, a less complex task than our extrapolation, and employed mean absolute percentage error (MAPE) rather than our use of SMAPE – obtained MAPE between 10 and 59.4 (see Miah et al., 2023 for an overview of these studies). Our results also show that the most important data sources in our study are crowdsourced Strava flow data (in the area around the counting station), infrastructure data (use of the area, street level indicators, and points of interest), time indicators, and motorized traffic (city-wide). Furthermore, we find that the prediction error varies greatly between locations, which means that the model is able to predict certain streets very well and others much less so (with no apparent pattern). This is also anecdotally shown in the proof of concept, where the model performs well in capturing temporal trends and identifying high-volume areas, but shows shortcomings in reproducing intricate geographic nuances. The feature selection process results in some features derived from averaging over large radii, which means that the prediction is in part optimized towards performing well over large area, temporal variations. The model's ability to reproduce intricate geographic nuances could be improved by better data that both allow for creating more street-level features as well as having more ground truth counter locations. The combination of selected street-level/small-radius features with large-radii features produces the best possible predictions by balancing the need to capture broader trends and localized variations. Finally, collecting sample counts for unseen locations not only drastically reduces the error across all locations but also the variance across locations. The decrease is at around 2/3 on average. This experiment showed that spending resources to collect additional short-term counts may be worthwhile.

Experts tasked by municipal governments can replicate our model using data that are already owned by the city or can be obtained from third-party providers. Although implementing the coding process demands a certain level of expertise, the models themselves are computationally efficient, enabling practical application in a municipal setting (Appendix A.3). Counting stations provide bicycle counts at limited locations. To predict bicycle counts beyond existing municipal counting stations, we advise

obtaining additional data sources on regional flow data from crowdsourced applications (Strava), comprehensive infrastructure data, and, if available, city-wide motorized traffic. We also recommend conducting multiple one-day sample counts at locations of interest to obtain more accurate results. Each day of sampling leads to a significant improvement in the estimate for that location. Using 10 days of sample data, our model provides policymakers with accurate and reliable estimates. Such estimates can allow them to make evidence-based decisions about infrastructure improvements or repairs. Busier roads can be prioritized, and financial expenditures can be justified by the number of cyclists they may benefit. Similarly, civil society can use such estimates to advocate for local infrastructure improvement needs.

In future research, more complex modeling approaches that take into account spatial and temporal dependencies can be another promising direction. Such approaches may also benefit from more ground truth data, particularly from more continuous counting stations to cover spatial variability. We hypothesize that more ground truth data could further improve predictions for unseen locations and possibly reduce the need for sample data collection. Expanding the case study of Berlin to a comparative analysis with other cities could shed light on the generalizability of the approach. Finally, while this article does not grapple with how cycling volumes are practically used, future studies could explore the applicability for planning, policy-making, and urban traffic modeling techniques.

4. Methods

4.1. Data description

A table explaining each individual feature is included in [Appendix A.4, Table 6](#). In the following, we elaborate on the data sources. All datasets are publicly available via the sources cited with one exception. Strava Metro provides its crowdsourced app data upon request (Strava Metro, 2023).

4.1.1. Bicycle counting stations data

The Berlin city administration collects data on bicycle volume at various locations (Senate Department for the Environment, Mobility, Consumer and Climate Protection Berlin, 2023). The data come from long-term counting stations, which are permanent devices that identify passing bicycles through an electromagnetic field embedded in the ground. The city installed its first of 30 counting stations in 2012 and the most recent one in 2022. Ten of the stations are located on opposite sides of the street and also record the direction of flow, while in the other locations, there is only one counter for both directions. We sum counters on opposite sides of the same street into one count as we are interested in the number of bicycles passing by a certain location rather than their direction of flow. This reduces the number of counting locations to 20. Occasionally, counting stations are out of service (e.g., due to construction or malfunctioning), resulting in missing observations. We also exclude observations that are interpolated by the municipality, as the city does provide information on their interpolation method. Short-term bicycle counts have been conducted at 21 fixed locations repetitively on different dates by the city since 1983. We exclude short-term counters consisting of only one observation (1 day) from the analysis. Consequently, the data set comprises information from a total of 12 short-term locations. A map of the counting stations, a table detailing the number of observations per station as well as some basic descriptions of the measurements are included in the [Appendix A.5](#). There is no publicly available information on the criteria used by the city to determine the placement of these counting stations. Upon closer examination, we find that these stations tend to be located closer to the city center (and thus, also closer to, for example, shops and education centers), along high-traffic roads (with a speed limit of 50 km/h), and in streets with some sort of cycling lane (see for more details [Appendix A.6](#)). We recognize this selection bias. Literature concerned with the Traffic Sensor Location Problem (Owais, 2022) addresses the concept of placing sensors optimally given various purposes. However, to the best of our knowledge, there is currently no research comparing the optimal placement of bicycle sensors with their actual locations. Further research is needed.

4.1.2. *Bike-sharing data*

Bike sharing data is an emerging data source that is considered for monitoring cycling (Lee and Sener, 2020). As bike-sharing users are part of the cycling population, we hypothesize that bike-sharing data can serve as an indicator for predicting overall cycling volumes. To our knowledge, only one study (Miah et al., 2023) has examined the potential of bike-sharing data for this purpose. In this study, we build on this approach by exploring the predictive value of bike-sharing data in a new urban context. The bike-sharing data used in this study consists of individual trips from free-floating bike-share systems. In these systems, bicycles are available for pick up and return anywhere within the city, unlike systems dependent on designated stations for both pickup and drop-off. It comprises two distinct periods. The months from April to December 2019, covering the providers Call-a-bike and Nextbike, as provided to us by CityLab Berlin, 2019. And from June to December 2022, covering only Nextbike, which we collected ourselves via their application programmable interface (API) (Nextbike, 2020). The data is made available for download (Kaiser, 2023). These data provide details of individual trips, unlike crowdsourced data (Strava), which only offers aggregate counts. The 2022 bike-sharing data was collected as follows. Nextbike's application programming interface furnishes real-time information on the location of all accessible bicycles, each identified by a unique bike ID, at a minute-by-minute granularity. When a bike is rented, it is temporarily removed from the available list and reappears when it is returned. By querying the list at one-minute intervals, we can accurately record trips to the minute, providing precise departure and arrival points and the respective times for every trip. For both bike-sharing datasets, 2019 and 2022, only the start and end points of each trip are available. We impute the route using OpenStreetMap as of July 2022 and the designated routing algorithm tailored for bicycles. OpenStreetMaps facilitates route planning for different modes of transportation, by adapting the suggested route according to the chosen mode (OpenStreetMap contributors, 2017). It is important to note that the resulting trajectories are approximations of the actual routes taken by bike-sharing users. Based on the routed trips, we perform data cleaning to account for possible incorrectly recorded journeys. We exclude trips shorter than 100 m (0.64% of total trips), which may be due to errors in GPS measurements or aborted trips, for example, due to a broken bike. Similarly, we exclude trips longer than the 45 km diameter of Berlin (0.005%), shorter than 120 seconds (1.63%), and longer than 10 hours (6.05%), assuming that incorrect use of the rental system is the cause. Finally, we exclude trips with an average speed slower than 2 km/h (16.57%) or faster than 40 km/h (10.87%). After cleaning, the data contain 1,333,737 bike-sharing trips. We engineer the bike-sharing data based on the methodology proposed by Miah et al. (2023). For each counter, we count the number of bikes passing, the number of bikes whose rental started, and the number of bikes whose rentals ended within a certain radius within a day. Unlike their approach, which considers a single radius of 0.125 miles, we examine multiple radii: 100, 200 m, 500 m, 1000 m, 2000 m, 5000 m, 6000 m, and the entire city. While smaller radii are selected more often, we still include larger radii because they can play an important role, possibly because they have a more regional impact (Hankey and Lindsey, 2016). We do not create a feature on the number of bike-sharing trips per street segment, as the exact route between the start and end points is interpolated. We provide an example of feature engineering for the bike-sharing data in [Appendix A.7](#). The following limitations to the bike-sharing data remain. In Berlin, the bike-sharing market is diverse, with multiple providers. We were only able to acquire data from two providers. These two maintain a fleet of traditional bicycles, unlike other companies that primarily offer e-bikes. In addition, neither Nextbike nor Call-a-bike responded to requests for information or provided detailed information about their data. This lack of transparency raises concerns about the stability of bike IDs, potentially leading to the inclusion of fictitious rides in our dataset due to how we compute trips. Also, these data are potentially biased, as bike-sharing users differ from private bicycle users. We do not have demographic information about the users of Nextbike or Call-a-bike, but bike-sharing users tend to have higher incomes and education (Fishman, 2016).

4.1.3. *Crowdsourced app data*

We obtain crowdsourced app data from the Strava smartphone app, which allows users to track their speed, altitude gain, and exact route choice covered during physical activities such as cycling (Strava

Metro, 2023). Similar to bike sharing, Strava users represent a small and varying share of the cycling population (Conrow et al., 2018). However, usage patterns differ: Strava users tend to ride at about twice the speed of bike-share users, while bike-sharing is more prevalent at night, and Strava sees higher usage in the evening (see [Appendix A.9](#)). Previous research has demonstrated that incorporating crowdsourced data as features within an ML model, alongside additional data, can significantly enhance model performance in predicting bicycle volume (Kwigizile, Morgan Kwayu and Oh, 2022). Strava provides these city-specific data at no cost to organizations involved in designing, overseeing, or maintaining cycling infrastructure upon request (Strava Metro, 2023); however, access to these data is contingent on Strava's approval. While they typically grant access to city administrations, this could be subject to change in the future. Strava Metro has modified the data to protect individual privacy. The data includes only publicly available trip records (as opposed to trips taken in a private mode). They also do not provide individual trip information but instead, aggregate the trip counts into two formats. In the first format, various features are available at the "street segment" level, which covers a street between two intersections. The data is available on an hourly basis. In the second format, features are available for regions in the form of hexagons, each spanning approximately 0.66 km². In both formats, bicycle counts are rounded to the nearest multiple of five, for example, when seven cyclists pass a street segment, Strava rounds it to five. We use both the segment and the hexagon formats. For the segment data, the features include information on the number of journeys, whether the trip was made with an e-bike, the gender and age group of the user, and the average speed. We calculate the average of the available features for all segments within a certain radius of the counting station (as for the other features, 100 m, 200 m, 500 m, 1000 m, 2000 m, 5000 m, 6000 m, whole city) at the daily level. Additionally, we compute the features on the level of individual street segments. For the hexagon data, the features include the number of trips, the purpose of the trip (leisure or commute), and the time of the day (morning, midday, evening). We use the features for both the hexagon where the counter is located and the mean of the six surrounding hexagons. We include a wide range of features across both segment and hexagon data formats to allow the model to capture nuanced variations in cycling patterns that may influence overall cycling volume. With feature selection, the model identifies the most relevant predictors from this set. A graphical representation of our feature engineering process is available in [Appendix A.8](#), with all features listed in [Appendix A.4](#). The Strava data is based only on the voluntary recording of Strava app users, and it has a sampling bias in its user base (Lee and Sener, 2021). Based on the few demographic indicators included in the data, a sampling bias towards young males with an ambitious riding style is apparent: 89.57% of trips were recorded by male users, only 3.9% of the trips were recorded by users aged 55 and over, e-bike trips contribute only 0.17%, and the average speed is relatively high at 21.14 km/h (see [Appendix A.9](#) for more details). To check for a potential spatial bias, we compared the distribution of the Strava features at the counting stations to those on all street segments in Berlin (see descriptive statistics in [Appendix A.10](#)). Our findings show that the Strava usage at the counting stations generally reflects overall citywide patterns. This is true, especially for the flow characteristics. The only exception is that a larger portion of street segments across the city show zero Strava rides (i.e., no rides were recorded on these segments). However, there are apparent differences in the hexagon-based characteristics: data from the counting stations tends to have much higher values. Since traffic flow data is the primary predictor for our model, while hexagon data plays a lesser role (see [Figure 2a](#)), we consider this potential bias to have minimal impact.

4.1.4. Weather data

Cyclists are more exposed to environmental conditions than motorized traffic users, which affects their comfort while cycling and, consequently, their decision to use a bicycle. Research shows that weather conditions can have both positive and negative effects on cycling, resulting from both immediate and delayed weather effects (Miranda-Moreno and Nosal, 2011). The data we use comes from the German Weather Service and is provided by Meteostat (2022). We include various features at daily granularity. The weather indicators are for all of Berlin, that is, they are the same for all counting locations but vary over time.

4.1.5. Infrastructure data

We include infrastructure data on road conditions, points of interest, and land use around the counting stations. Cycling and road infrastructure play a critical role in increasing cyclists' perception of safety and, consequently, influence bicycle use (Møller and Hels, 2008). The infrastructure features also affect how many trips may be made to an area. Similarly, points of interest, such as schools and shops, can influence bicycle volumes at different times throughout the day and week (Strauss and Miranda-Moreno, 2013). From OpenStreetMap contributors, 2017, we obtained information about the maximum speed allowed for motorized traffic, the type of bike lane at the exact location of the counting station, and the number of different points of interest within different radii (as for the other features: 100, 200 m, 500 m, 1000 m, 2000 m, 5000 m, and 6000 m). In contrast to the radii used for the other data sources, we do not compute the features for the entire city. Doing so would result in constant features over time and space, providing no valuable information to our model. We also compute the distance of the counting stations from the city center, following the definition of a city center used by OpenStreetMap. Data from the city of Berlin provides information on land use, such as for parks or industry, which can impact the timing and volume of human frequenting in various areas. This data is collected at the "planning area" level: For urban planning purposes, the city is divided into planning areas that represent neighborhoods; each planning area has an average size of about 2 km². The city collected the indicators in 2015 (Berlin Open Data, 2022a; Senate Department for Urban Development, Building and Housing, 2023). We use data from the planning area surrounding each counting station. To standardize the measurements, we convert the features from square kilometers into percentages. For example, instead of stating that 0.5km² within the planning area surrounding the counting station is occupied by parks, we express it as 25% occupied by parks.

4.1.6. Centrality measures

We include network connectivity measures as features because they capture the importance of individual links, that is, roads, within the transport network – a factor associated with cycling behavior (Schön, Heinen, and Manum, 2024; Hochmair, Bardin, and Ahmouda, 2019; Lu et al., 2017). Using the Berlin road network (Berlin Open Data, 2024), we compute the following graph measures: degree, which reflects the number of direct connections a road segment has; betweenness, which indicates the frequency with which a segment serves as a bridge along the shortest paths in the network; closeness, which measures the average distance from a segment to all other segments, thus reflecting its accessibility; and the clustering coefficient, which assesses the degree to which a segment's neighboring links are interconnected. Formal definitions of these measures can be found in [Appendix A.11](#).

4.1.7. Socioeconomic data

Bicycle use varies with regard to socioeconomic characteristics such as age and gender (Goel et al., 2022). We obtain socioeconomic data from the city of Berlin (Berlin-Brandenburg Office of Statistics, 2023). We include socioeconomic features at the level of "planning areas" (see the infrastructure data for details). For each counter, we use the indicators of the planning area in which it is located. Since the socioeconomic data are only available until 2020, we use the 2019 observations for the same year and the 2022 observations for 2020. Therefore, the data has spatial and temporal variation, but the data for 2022 remains an approximation.

4.1.8. Motorized traffic data

Similar to the bicycle counting stations, the city administration conducts counts of motorized traffic at 267 counting stations (for 2019 and 2022) (Berlin Open Data, 2022b). To the best of our knowledge, we are unaware of any studies that have attempted to predict bicycle volume from motorized traffic counts. We hypothesized that this could be a valuable source of additional information as bicycle counts and motorized traffic counts may show similar patterns in terms of commuting peaks, weekday/weekend behavior, and locations of interest. The data includes the volume, type, and speed of motorized vehicles collected at various locations in Berlin. The detectors measure the features only on one side of the road (e.g., only eastbound traffic). We compute the respective mean values of these motorized traffic features of

all traffic counters within a 6-kilometer radius and for the city as a whole, all on a daily basis. The choice of a 6-kilometer radius as our spatial unit is intentional, as it is the smallest possible radius for the feature to be available for each counting station. The main drawback of the data is that the motorized traffic counting stations are unevenly distributed throughout the city (see [Appendix A.12](#)). Therefore, not only do we have to employ a large radius, but we also have to compute the features for each counting station based on a different number of motorized traffic counting stations.

4.1.9. *Holiday and time data*

Traffic data can exhibit strong seasonality. We, therefore, include several time indicators: the day of the week, the day of the month, the month itself as features, the year, and whether it is a weekday. We also use features that indicate the presence of each public and school holiday (Senate Department for Education, Youth and Family, 2022).

4.1.10. *Pre-processing of the combined data*

In the merged dataset, an observation represents a daily measurement from one counting station. Additionally, we exclude any features that are constant across all observations. This is the case for the number of hospitals within 100, 200, and 500 m, the number of industries for all radii, and the percentage of land used for horticulture, as they are all zero. The socioeconomic data is also missing for one counting station, which we replace with the mean of the respective features across all other counting stations. Based on this preprocessed data, we conduct feature selection as detailed in the next section.

4.2. *Data analysis*

4.2.1. *Models and algorithms*

We implement all models with the Python library scikit-learn (Pedregosa et al., 2011). Based on the results for extrapolating daily and annual street-level bicycle volumes, we selected Extreme Gradient Boosting (XGBoost) for further analysis of the feature importance and the sample data collection experiment, as it demonstrated slightly better performance in our comparative tests. XGBoost is an ML algorithm that combines boosting and regularisation techniques. By iteratively adding decision trees to an ensemble model, it corrects the errors of the previous trees, resulting in a more robust and accurate model compared to standard decision trees or random forests. The trees are trained using a gradient descent optimization method, which updates the weights of the features to minimize a given loss function. In addition, the algorithm uses a technique called tree pruning to remove unnecessary leaves and nodes from the trees. XGBoost also incorporates regularization techniques, which help mitigate multicollinearity by reducing the influence of correlated or redundant features, ensuring that the model remains stable. For information on the other models, we would like to refer the reader to Géron (2022).

4.2.2. *Addressing the skewed distribution of the counting stations' measurements*

The target variable, the measurements from counting stations, follows a non-normal distribution characterized by a pronounced right skew (see [Appendix A.5](#)). To account for this, we transformed our target variable using a logarithmic transformation (see [Appendix A.13](#) for QQ plot). This approach reduces the impact of extreme values on our models and ensures that the errors are more robustly distributed across different types of roads, including those with low traffic volumes. We used the log-transformed target variable during model training and applied the inverse transformation before calculating test errors to provide interpretable results on the original scale.

4.2.3. *Hyperparameter tuning*

For XGBoost, we tune the following hyperparameters with random search: the learning rate (controls the step size during the optimization process), the maximum depth of each tree (deeper trees can capture more complex relationships but can cause overfitting), the fraction of features used when constructing each tree (reducing overfitting also introducing randomness), the minimum sum of instance weight needed in a

child (can prevent overfitting by controlling the minimum amount of instances required in each leaf) and a regularization parameter that encourages pruning of the tree (Brownlee, 2016). For the hyperparameters of the other models, we would like to refer the reader to [Appendix A.14](#).

4.2.4. Feature selection

Given the richness of our feature set, we implement model-specific feature selection (FS) techniques to minimize computational demands and enhance model performance. For each model, we evaluated several methods, including Select KBest, recursive feature elimination with CV, and sequential FS. Ultimately, we employ recursive feature elimination for linear regression, gradient boosting, random forest, sequential feature selection for XGBoost and decision tree, while applying Select KBest for SVM and SNN (Pedregosa et al., 2011) on both the 0–24 h and 7–19 h datasets. For the XGBoost model, we observe that none of the centrality measures or socioeconomic features were selected. Instead, the FS prioritized several street-level and small-radius features (e.g., 100 m, 200 m) from Strava, infrastructure, and bike-sharing data, along with a few larger-radius features (i.e., 1000 m, 2000 m) ([Appendix Table 4](#)). For the other models, we find much variation in the selected features, with 243 of the 257 features selected at least once. For all models, features based on smaller radii tend to be selected more often, but also, for almost every model, a feature based on a large radii is included. We find Strava (flow in the area) and motorized (city-wide) indicators selected relatively often across models (see [Appendix A.15](#) for an overview and the number of selected features).

4.2.5. Error metrics

For benchmarking, we choose two error metrics, MAE and SMAPE, which are defined as follows, with n the number of observations, y_i the true value, and \hat{y}_i the prediction of the variable of interest:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4.1)$$

$$\text{SMAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}. \quad (4.2)$$

We have chosen these metrics over the more commonly used error metrics: root mean squared error (RMSE), mean absolute percentage error (MAPE), or standard errors based on the underlying distribution. The counting station measurements include several high values. Compared to the RMSE, the MAE places less emphasis on extreme values, which is more suited to the right tail distribution. See [Appendix A.5](#) for histograms and boxplots of the counts. We have chosen SMAPE over MAPE to measure the relative error, as it yields small percentage errors when the true value is very high. Additionally, SMAPE gives a symmetric measure that considers both overestimation and underestimation errors equally.

4.2.6. Grouped feature importance

Computing feature importance for groups of features cannot be simply done by summing individual feature importance scores. Neither can one sum individual feature importance for tree-based methods. This method often overfits, boosting features that contribute little, and thus, they cannot be summed at the group level since this would overweight groups with many features (Loughrey and Cunningham, 2005; Breiman, 2001). Nor can one sum up permutation-based features, as all features besides the one in question are known during the permutation, which does not sufficiently reveal the impact of a particular feature group when summed up (Plagwitz et al., 2022).

Here, we use the grouped feature importance as introduced by Plagwitz et al., 2022. The data is split in the sense of cross-validation into training and test data. On each fold, the following is computed: A model is trained on the training data. The test data is replicated a certain amount of times, and in each replication, the features belonging to a feature group in question are permuted. The model is then applied to the permuted test sets. The change in performance is estimated and averaged across all permuted test sets. This process is repeated for every feature group. The mean between the cross-validation folds returns the final grouped feature importance score, which provides the information gain per group. These scores are not comparable across models but only within each model.

4.2.7. Sample weights in training

We employ sample weights during model training with the sample data to enhance the model's emphasis on these particular observations. Sample weights assign different weights to individual observations in the training dataset. This is useful when dealing with imbalanced datasets or when certain samples are more critical than others. The latter is the case in our setting. In XGBoost, sample weights can be assigned to each instance, influencing the contribution of that instance's error to the overall loss function during training. This way, samples with higher weights contribute more to the model's updates, thus affecting the model's learning process (Pedregosa et al., 2011).

Abbreviations

AADB	average annual daily bicycle volume
CV	cross validation
FS	feature selection
GPI	grouped permutation Importance
LOGO	leave-one-group-out
MAE	mean absolute error
MAPE	mean absolute percentage error
ML	machine learning
RMSE	root mean squared error
SMAPE	symmetric mean absolute percentage error

Open peer review. To view the open peer review materials for this article, please visit <http://doi.org/10.1017/eds.2025.5>.

Acknowledgements. We thank E. Kolibacz for his technical support. We are grateful to CityLab Berlin for providing their bike-sharing data.

Author contribution. Conceptualization & Methodology: S.K., L.K.; Formal analysis, Investigation: S.K.; Resources: S.K., L.K., N.K.; Writing - Original Draft: S.K., L.K.; Writing - Review & Editing: S.K., L.K., N.K. All authors approved the final submitted draft.

Competing interest. The authors declare no competing interests exist.

Data availability statement. The preprocessed data has been made available on zenodo: <https://doi.org/10.5281/zenodo.14499983>. Notably, the Strava features within the dataset have been modified due to the non-permissibility of public dissemination of this particular data.

Funding statement. This research was supported by grants from the European Union's Horizon Europe research and innovation program under Grant Agreement No 101057131, Climate Action To Advance HealthY Societies in Europe (CATALYSE). Furthermore, the authors acknowledge support through the Emmy Noether grant KL 3037/1-1 of the German Research Foundation (DFG). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethical statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country (Germany).

References

- Berlin Open Data** (2022a) *Nutzung Stadtstruktur 2015 [Use of Urban Structure 2015]*. <https://daten.berlin.de/kategorie/geographie-und-stadtplanung>.
- Berlin Open Data** (2022b) *Verkehrsdetektion Berlin [Traffic Detection Berlin]*. <https://daten.berlin.de/datensaetze/verkehrsdetektion-berlin>.
- Berlin Open Data** (2024) *Detailnetz Berlin [Detailed Network Berlin]*. <https://fbinter.stadt-berlin.de/fb/index.jsp>.
- Berlin-Brandenburg Office of Statistics** (2023) *Kommunalatlas Berlin [Municipal Atlas Berlin]*. <https://instantatlas.statistikberlin-brandenburg.de/instantatlas/interaktivekarten/kommunalatlas/atlas.html>.
- Breiman L** (2001) Random forests. *Machine Learning* 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Brownlee J** (2016, August) *XGBoost with Python: Gradient Boosted Trees with XGBoost and Scikit-Learn*. Google-Books-ID: HgmqDwAAQBAJ. Machine Learning Mastery.
- CityLab Berlin** (2019) *Shared Mobility Flows*. <https://github.com/technologiestiftung/bike-sharing> (visited on 03/01/2022).

- Conrow L, Wentz E, Nelson T and Pettit C** (2018) Comparing spatial patterns of crowdsourced and conventional bicycling datasets. *Applied Geography* 92, 21–30. <https://doi.org/10.1016/j.apgeog.2018.01.009>. <https://linkinghub.elsevier.com/retrieve/pii/S0143622817310548>.
- Courty B, Schmidt V, Luccioni S, Goyal-Kamal, MarionCoutarel, Feld B, Lecourt J, LiamConnell, Saboni A, Inimaz, supatomic, Léval M, Blanche L, Cruveiller A, ouminasara, Zhao F, Joshi A, Bogroff A, de Lavoreille H, Laskaris N, Abati E, Blank D, Wang Z, Catovic A, Alencon M, Stęchly M, Bauer C, de Araújo LON, JPW and MinervaBooks** (2024, May). *mlco2/codecarbon: v2.4.1*. <https://doi.org/10.5281/zenodo.11171501>.
- Dadashova B and Griffin GP** (2020) Random parameter models for estimating statewide daily bicycle counts using crowdsourced data. *Transportation Research Part D: Transport and Environment* 84, 102368. <https://doi.org/10.1016/j.trd.2020.102368>.
- Das S and Tsapakis I** (2020) Interpretable machine learning approach in estimating traffic volume on low-volume roadways. *International Journal of Transportation Science and Technology* 9(1), 76–88. <https://doi.org/10.1016/j.ijtst.2019.09.004>.
- Dill J** (2009) Bicycling for transportation and health: The role of infrastructure. *Journal of Public Health Policy* 30(S1), S95–S110. <https://doi.org/10.1057/jphp.2008.56>.
- El Esawey M** (2018) Daily bicycle traffic volume estimation: Comparison of historical average and count models. *Journal of Urban Planning and Development* 144(2), 04018011. <https://doi.org/10.3141/2443-12>.
- European Metropolitan Transport Authorities** (2023) *Barometer 2022 – Based on 2013–2023 data, published June 2024*. <https://www.emta.com/publications/article-emta-barometer-of-public-transport/>.
- Fishman E** (2016) Bikeshare: A review of recent literature. *Transport Reviews* 36(1), 92–113. <https://doi.org/10.1080/01441647.2015.1033036>.
- Garrard J, Rose G and Lo SK** (2008) Promoting transportation cycling for women: The role of bicycle infrastructure. *Preventive Medicine* 46(1), 55–59. Elsevier. <https://doi.org/10.1016/j.ypmed.2007.07.010>.
- Géron A** (2022) *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.
- Goel R, Goodman A, Aldred R, Nakamura R, Tatah L, Garcia LMT, Zapata-Diomedí B, de Sa TH, Tiwari G, de Nazelle A, Tainio M, Buehler R, Götschi T, Woodcock J** (2022) Cycling behavior in 17 countries across 6 continents: Levels of cycling, who cycles, for what purpose, and how far? *Transport Reviews* 42(1), 58–81. <https://doi.org/10.1080/01441647.2021.1915898>.
- Griswold JB, Medury A and Schneider RJ** (2011) Pilot models for estimating bicycle intersection volumes. *Transportation Research Record* 2247.1, 1–7. <https://doi.org/10.3141/2247-01>.
- Hankey S and Lindsey G** (2016) Facility-demand models of peak period pedestrian and bicycle traffic: Comparison of fully specified and reduced-form models. *Transportation Research Record* 2586(1), pp. 48–58. <https://doi.org/10.3141/2586-06>.
- Hankey S, Lindsey G and Marshall J** (2014) Day-of-year scaling factors and design considerations for nonmotorized traffic monitoring programs. *Transportation Research Record: Journal of the Transportation Research Board* 2468(1), 64–73. <https://doi.org/10.3141/2468-08>.
- Hochmair HH, Bardin E and Ahmouda A** (2019) Estimating bicycle trip volume for Miami-Dade county from Strava tracking data. *Journal of Transport Geography* 75, 58–69. <https://doi.org/10.1016/j.jtrangeo.2019.01.013>.
- Holmgren J, Aspegren S and Dahlströma J** (2017) Prediction of bicycle counter data using regression. *Procedia Computer Science* 113, 502–507. <https://doi.org/10.1016/j.procs.2017.08.312>.
- Kaiser SK** (2023) *Bike-Sharing Data Berlin from Nextbike and Call-a-Bike for 2019 and 2022*. <https://doi.org/10.5281/zenodo.10046530>.
- Kwigizile V, Kwayu KM and Oh J-S** (2022) Leveraging the spatial-temporal resolution of crowdsourced cycling data to improve the estimation of hourly bicycle volume. *Transportation Research Interdisciplinary Perspectives* 14, 100596. <https://doi.org/10.1016/j.trip.2022.100596>.
- Larsen J, Patterson Z and El-Geneidy A** (2013) Build it. But where? The use of geographic information Systems in identifying locations for new cycling infrastructure. *International Journal of Sustainable Transportation* 7(4), 299–317. <https://doi.org/10.1080/15568318.2011.631098>.
- Lee K and Sener IN** (2020) Emerging data for pedestrian and bicycle monitoring: Sources and applications. *Transportation Research Interdisciplinary Perspectives* 4, 100095. <https://doi.org/10.1016/j.trip.2020.100095>.
- Lee K and Sener IN** (2021) Strava Metro data for bicycle monitoring: a literature review. *Transport Reviews* 41(1), 27–47. <https://doi.org/10.1080/01441647.2020.1798558>.
- Loughrey J and Cunningham P** (2005) Overfitting in wrapper-based feature subset selection: the harder you try the worse it gets. In Bramer M, Coenen F and Allen T (eds.), *Research and Development in Intelligent Systems XXI*. London: Springer, pp. 33–43. https://doi.org/10.1007/1-84628-102-4_3.
- Lu T, Buehler R, Mondschein A and Hankey S** (2017) Designing a bicycle and pedestrian traffic monitoring program to estimate annual average daily traffic in a small rural college town. *Transportation Research Part D: Transport and Environment* 53, 193–204. <https://doi.org/10.1016/j.trd.2017.04.017>.
- meteostat** (2022) *The Weather's Record Keeper*. <https://meteostat.net/en/>.
- Miah MM, Hyun KK, Mattingly SP, Broach J, McNeil N and Kothuri S** (2022) Challenges and opportunities of emerging data sources to estimate network-wide bike counts. *Journal of Transportation Engineering, Part A: Systems* 148(3), pp. 04021122. <https://doi.org/10.1061/JTEPBS.0000634>.
- Miah MM, Hyun KK, Mattingly SP and Khan H** (2023) Estimation of daily bicycle traffic using machine and deep learning techniques. *Transportation*, 50(5), 1631–1684. <https://doi.org/10.1007/s11116-022-10290-z>.

- Miranda-Moreno LF and Nosal T** (2011) Weather or not to cycle: Temporal trends and impact of weather on cycling in an urban environment. *Transportation Research Record* 2247(1), 42–52. <https://doi.org/10.3141/2247-06>.
- Møller M and Hels T** (2008) Cyclists' perception of risk in roundabouts. *Accident Analysis & Prevention* 40(3), 1055–1062. <https://doi.org/10.1016/j.aap.2007.10.013>.
- Morrison CN, Thompson J, Kondo MC and Beck B** (2019) On-road bicycle lane types, roadway characteristics, and risks for bicycle crashes. *Accident Analysis & Prevention* 123, 123–131. <https://doi.org/10.1016/j.aap.2018.11.017>.
- Nelson T, Roy A, Ferster C, Fischer J, Brum-Bastos V, Laberee K, Yu H, and Winters M** (2021) Generalized model for mapping bicycle ridership with crowdsourced data. *Transportation Research Part C: Emerging Technologies* 125, 102981. <https://doi.org/10.1016/j.trc.2021.102981>.
- Nextbike** (2020) *Official Nextbike API Documentation*. <https://github.com/nextbike/api-doc>.
- Oja P, Titze S, Bauman A, de Geus B, Krenn P, Reger-Nash B, and Kohlberger T** (2011) Health benefits of cycling: a systematic review. *Scandinavian Journal of Medicine & Science in Sports* 21(4), 496–509. <https://doi.org/10.1111/j.1600-0838.2011.01299.x>.
- Olmos LE, Olmos LE, Tadeo MS, Vlachogiannis D, Alhasoun F, Espinet Alegre X, Ochoa C, Targa F, and González MC** (2020) A data science framework for planning the growth of bicycle infrastructures. *Transportation Research Part C: Emerging Technologies* 115, 102640. <https://doi.org/10.1016/j.trc.2020.102640>.
- OpenStreetMap contributors** (2017) *Planet Dump*. Retrieved from <https://planet.osm.org>. Accessed in July 2023. <https://www.openstreetmap.org>.
- Owais M** (2022) Traffic sensor location problem: Three decades of research. *Expert Systems with Applications* 208, 118134 <https://doi.org/10.1016/j.eswa.2022.118134>.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E** (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12, 2825–2830. (visited on 04/19/2023)
- Plagwitz L, Brenner A, Fujarski M, Varghese J** (2022) Supporting AI-explainability by analyzing feature subsets in a machine learning model. *Challenges of Trustable AI and Added-Value on Health*, 109–113. <https://doi.org/10.3233/SHTI220406>.
- Pörtner H-O, Roberts DC, Poloczanska ES, Mintenbeck K, Tignor M, Alegría A, Craig M, Langsdorf S, Lösschke S, Möller V, and Okem A** (2022) IPCC, 2022: Summary for policymakers. In Shukla PR, Skea J, Slade R, Al Khourdajie A, van Diemen R, McCollum D, Pathak M, Some S, Vyas P, Fradera R, Belkacemi M, Hasija A, Lisboa G, Luz S, Malley J (eds.) *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge/New York: Cambridge University Press.
- Romanillos G, Zaltz Austwick M, Ettema D and De Kruijff J** (2016) Big data and cycling. *Transport Reviews* 36(1), 114–133. <https://doi.org/10.1080/01441647.2015.1084067>.
- Roy A, Nelson TA, Fotheringham AS and Winters M** (2019) Correcting bias in crowdsourced data to map bicycle ridership of all bicyclists. *Urban Science* 3(2), 62. <https://doi.org/10.3390/urbansci3020062>.
- Sanders RL, Frackelton A, Gardner S, Schneider R and Hintze M** (2017) Ballpark method for estimating pedestrian and bicyclist exposure in Seattle, Washington: Potential option for resource-constrained cities in an age of big data. *Transportation Research Record: Journal of the Transportation Research Board* 2605(1), 32–44. <https://doi.org/10.3141/2605-03>.
- Schön P, Heinen E and Manum B** (2024) A scoping review on cycling network connectivity and its effects on cycling. *Transport Reviews* 44(4), 912–936. <https://doi.org/10.1080/01441647.2024.2337880>.
- Sekula P, Marković N, Laan ZV and Sadabadi KF** (2018) Estimating historical hourly traffic volumes via machine learning and vehicle probe data: A Maryland case study *Transportation Research Part C: Emerging Technologies* 97, 147–158. <https://doi.org/10.1016/j.trc.2018.10.012>.
- Senate Department for Education, Youth and Family** (2022) *Ferientermine [Vacation Dates]*. <https://www.berlin.de/sen/bjf/service/kalender/ferien/artikel.420979.php>.
- Senate Department for the Environment, Mobility, Consumer and Climate Protection Berlin** (2022) *Jahresdatei geprüfter Rohdaten der Radzählstellen [Annual File of Audited Raw Data of Bicycle Counting Stations]*. <https://www.berlin.de/sen/uvk/verkehr/verkehrsplanung/radverkehr/weitere-radinfrastruktur/zaehlstellen-und-fahrradbarometer/>.
- Senate Department for the Environment, Mobility, Consumer and Climate Protection Berlin** (2023) *Zählstellen und Fahrradbarometer: Fahrradverkehr in Zahlen [Counting Stations and Bicycle Barometer: Bicycle Traffic in Figures]*. <https://www.berlin.de/sen/uvk/verkehr/verkehrsplanung/radverkehr/weitere-radinfrastruktur/zaehlstellen-und-fahrradbarometer/>.
- Senate Department for Urban Development, Building and Housing** (2023) *Lebensweltlich orientierte Räume (LOR) Berlin [Living Environment Oriented Rooms (LOR) in Berlin] Shapefiles*. <https://www.berlin.de/sen/sbw/stadt/stadtwissen/sozialraumorientierte-planungsgrundlagen/lebensweltlich-orientierte-raeume/>.
- Strauss P and Miranda-Moreno LF** (2013) Spatial modeling of bicycle activity at signalized intersections. *Journal of Transport and Land Use* 6(2), 47–58. <https://doi.org/10.5198/jtlu.v6i2.296>.
- Strava Metro** (2023) *Strava Metro – Berlin Data*. <https://metro.strava.com/>.
- Woodcock J, Edwards P, Tonne C, Armstrong BG, Ashiru O, Banister D, Beevers S, Chalabi Z, Chowdhury Z, Cohen A, Franco OH, Haines A, Hickman R, Lindsay G, Mittal I, Mohan D, Tiwari G, Woodward A, and Roberts I** (2009) Public health benefits of strategies to reduce greenhouse-gas emissions: Urban land transport. *The Lancet* 374(9705), 1930–1943. [https://doi.org/10.1016/S0140-6736\(09\)61714-1](https://doi.org/10.1016/S0140-6736(09)61714-1).

- Wu X, Lindsey G, Fisher D and Wood SA** (2017) Photos, tweets, and trails: Are social media proxies for urban trail use? *Journal of Transport and Land Use* 10(1), 789–804. <https://doi.org/10.5198/jtlu.2017.1130>.
- Yi Z, Liu XC, Markovic N and Phillips J** (2021) Inferencing hourly traffic volume using data-driven machine learning and graph theory. *Computers, Environment and Urban Systems* 85, 101548. <https://doi.org/10.1016/j.compenvurbsys.2020.101548>.
- Zahedian S, Sekula P, Nohekhan A, and Vander Laan Z** (2020) Estimating hourly traffic volumes using artificial neural network with additional inputs from automatic traffic recorders. *Transportation Research Record* 2674(3), 272–282. <https://doi.org/10.1177/0361198120910737>.

A. Appendix

A.1. Features grouping for the feature importance

Table 4. Grouping of the features for the feature importance evaluation in Section 2.3: This table presents the features selected for the XGBoost model for 0–24 h (all day), grouped into their respective categories

Group	Features
Time	Weekday, weekend, month, year
Holidays	School holiday, public holiday
Weather	Maximum snow depth, wind direction
Bike-sharing (flow in area)	No. bikes rented (within 2000 m)
Bike-sharing (origin/dest.)	No. bikes originating (within 100 m), no bikes returned (within 100 m)
Infrastructure (street level)	Maximum speed, bicycle lane type
Infrastructure (points of interest)	No. of shops (within 500 m)
Infrastructure (use of area)	Percent of area used for farming, parks, private gardening and forest
Motorized (city-wide)	No. of cars
Motorized (area)	Average speed of lorries (within 6000 m)
Strava (origin/destination)	No. of trips originating by time of the day (weekend) in the respective hexagon
Strava (flow street level)	No. of rides (e-bikes only), no. of rides by various age groups (55–64 and 64+), and by sex (unspecified gender)
Strava (flow area)	No. of rides (e-bikes only within 100 m, 200 m, 500 m, 1000 m, and 2000 m), by gender (unspecified gender within 100 m, 200 m, and 2000 m; male within 500 m), by age group (55–64 and 65+ within 100 m and 200 m)

A.2. Graphic explanation: spatial extrapolation using additional sample count data

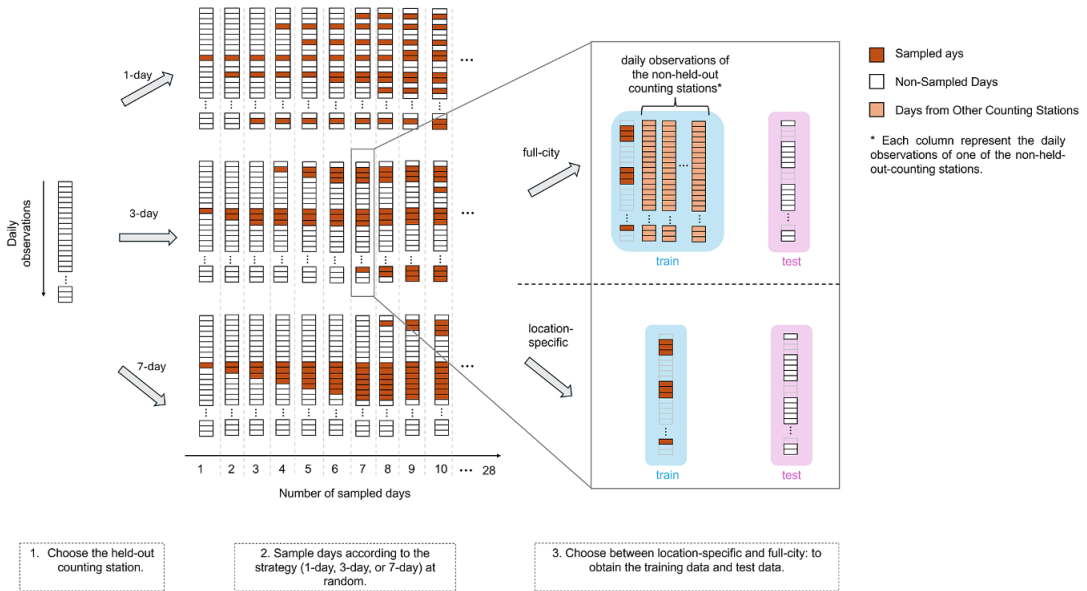


Figure 4. Workflow for estimating errors using sample counts in a LOGO evaluation. Each counting station is held out once, with the figure providing an example for one such counter. A sampling strategy – 1-day, 3-day, or 7-day – is selected, according to which up to 28 days from the held-out counting station are sampled. These days are the location specific training data (depicted in dark orange). The test data consists of unsampled observations from the same station (depicted in white). In the case of the location-specific model, we train the model on these data alone. For the city-wide model, the training data includes both the observations of the held-out station and data from other stations (depicted in light orange). For space considerations, the figure illustrates this process for the 3-day strategy with seven sampled days only.

A.3. Computation time and energy consumption

Table 5. Computation time and energy consumption for various tasks using XGBoost. The table reflects the performance of a 128-CPU server, illustrating the efficiency of the model in terms of both time and energy. Training and testing times are based on the task of predicting the short-term counting stations’ data (Table 3, column (3)). The energy consumption was computed with CodeCarbon (Courty et al., 2024)

Task	Computation time	Energy consumed for	
		RAM	all CPUs
Sequential feature selection	2532.53 s (≈ 42 min)	0.132732 kWh	0.079092 kWh
Hyperparameter tuning	50.78 s	0.002594 kWh	0.001546 kWh
Training	1.69 s	0.000021 kWh	0.000013 kWh
Testing	1.28 s	0.000001 kWh	0.000001 kWh

A.4. Detailed feature list

Table 6. Overview of all considered features

Feature name	Further explanation	Spatial scope	Timing for daily model	# Features	Type	Scaling	Data source
Time features							
Year	Measurement from 2019 or 2022	Stationary	Yearly	1	Binary		Inherent
Month	Indicating January through December	Stationary	Monthly	1	Numerical	One-hot-encoded	Inherent
Day of month	Indicating numerical day of month	Stationary	Monthly	1	Numerical	Standardized	Inherent
Weekday	Indicating Monday through Sunday	Stationary	Daily	1	Numerical	One-hot-encoded	Inherent
Weekend centrality measures	If Saturday or Sunday	Stationary	Daily	1	Binary		Inherent
Degree	See Appendix A.11 for formal explanation	Counting station location	Daily	1	Numerical	Standardized	Based on Berlin Open Data (2024)
Betweenness	see Appendix A.11 for formal explanation	Counting station location	Daily	1	Numerical	Standardized	Based on Berlin Open Data (2024)
Closeness	See Appendix A.11 for formal explanation	Counting station location	Daily	1	Numerical	Standardized	Based on Berlin Open Data (2024)
Clustering coefficient	See Appendix A.11 for formal explanation	Counting station location	Daily	1	Numerical	Standardized	Based on Berlin Open Data (2024)
Vacation and holiday features							
School holiday	Presence of school holiday	Stationary	Daily	1	Binary		Senate Department for Education, Youth and Family (2022)

Continued

Table 6. *Continued*

Feature name	Further explanation	Spatial scope	Timing for daily model	# Features	Type	Scaling	Data source
Public holiday	Presence of public holiday	Stationary	Daily	1	Binary		Senate Department for Education, Youth and Family (2022)
Bike-sharing Originating/returned/rented	number of	Within a certain radius (100 m, 200 m, 500 m, 1000 m, 2000 m, 5000 m, 6000 m) to the counter	Daily	21	Numerical	Standardized	Nextbike (2020) and web scraped from Nextbike, as well as Call-a-bike
Originating/returned/rented	number of	Within the whole city	Daily	3	Numerical	Standardized	Nextbike (2020) and web scraped from Nextbike, as well as Call-a-bike
Strava No. of trips (originating/arriving)	number of	In the respective hexagon	Daily	2	Numerical	Standardized	Strava Metro (2023)
No. of trips (overall/originating/arriving)	number of	In the six neighboring hexagons	Daily	2	Numerical	Standardized	Strava Metro (2023)
No. of trips (originating/arriving) by purpose (leisure/commute)	number of	In the respective hexagon	Daily	4	Numerical	Standardized	Strava Metro (2023)

Continued

Table 6. Continued

Feature name	Further explanation	Spatial scope	Timing for daily model	# Features	Type	Scaling	Data source
No. of trips (originating/ arriving) by purpose (leisure/ commute)	number of	In the six neighboring hexagons	Daily	4	Numerical	Standardized	Strava Metro (2023)
No. of trips (originating/ arriving) by time of the day (morning/ midday/ evening/ overnight/ weekday/ weekend)	number of	In the respective hexagon	Daily	12	Numerical	Standardized	Strava Metro (2023)
No. of trips (originating/ arriving) by time of the day (morning/ midday/ evening/ overnight/ weekday/ weekend)	number of	In the six neighboring hexagons	Daily	12	Numerical	Standardized	Strava Metro (2023)
No. of trips (all bikes/ e-bikes only)	number of	In the segments within a different radii (per street segment, 100 m, 200 m, 500 m, 1000 m, 2000 m, 5000 m, 6000 m)	Daily	16	Numerical	Standardized	Strava Metro (2023)

Continued

Table 6. *Continued*

Feature name	Further explanation	Spatial scope	Timing for daily model	# Features	Type	Scaling	Data source
No. of trips (all bikes/e-bikes only)	number of	In the whole city of Berlin	Daily	2	Numerical	Standardized	Strava Metro (2023)
No. of individuals	number of	In the segments within a different radii (per street segment, 100 m, 200 m, 500 m, 1000 m, 2000 m, 5000 m, 6000 m)	Daily	8	Numerical	Standardized	Strava Metro (2023)
No. of individuals	Number of	In the whole city of Berlin	Daily	1	Numerical	Standardized	Strava Metro (2023)
No. of rides by sex (female, male and unspecified gender)	number of	In the segments within a different radii (per street segment, 100 m, 200 m, 500 m, 1000 m, 2000 m, 5000 m, 6000 m)	Daily	24	Numerical	Standardized	Strava Metro (2023)
No. of rides by sex (female, male and unspecified gender)	Number of	In the whole city of Berlin	Daily	3	Numerical	Standardized	Strava Metro (2023)
No. of rides by various age groups (18–34, 35–54, 55–64 and 65+)	Number of	In the segments within a different radii (per street segment, 100 m, 200 m, 500 m, 1000 m, 2000 m, 5000 m, 6000 m)	Daily	32	Numerical	Standardized	Strava Metro (2023)

Continued

Table 6. Continued

Feature name	Further explanation	Spatial scope	Timing for daily model	# Features	Type	Scaling	Data source
No. of rides by various age groups (18–34, 35–54, 55–64 and 65+)	Number of	In the whole city of Berlin	Daily	4	Numerical	Standardized	Strava Metro (2023)
Average speed	Number of	In the segments within a different radii (per street segment, 100 m, 200 m, 500 m, 1000 m, 2000 m, 5000 m, 6000 m)	Daily	8	Numerical	Standardized	Strava Metro (2023)
Average speed	Number of	In the whole city of Berlin	Daily	1	Numerical	Standardized	Strava Metro (2023)
Infrastructure Latitude and longitude		Counting station location	Stationary	2	Numerical	Standardized	Senate Department for the Environment, Mobility, Consumer and Climate Protection Berlin (2022)
Distance to city center	In km	counting station location	Stationary	1	Numerical	Standardized	OpenStreetMap contributors (2017)
Maximum speed	In km/h	Counting station location	Stationary	1	Categorical	One-hot-encoded	OpenStreetMap contributors (2017)

Continued

Table 6. Continued

Feature name	Further explanation	Spatial scope	Timing for daily model	# Features	Type	Scaling	Data source
Bicycle lane type		Counting station location	Stationary	1	Categorical	One-hot-encoded	OpenStreetMap contributors (2017)
No. of shops		Within a certain radius (100 m, 200 m, 500 m, 1000 m, 2000 m, 5000 m, 6000 m)	Stationary	7	Numerical	Standardized	OpenStreetMap contributors (2017)
No. of education		Within a certain radius (100 m, 200 m, 500 m, 1000 m, 2000 m, 5000 m, 6000 m)	Stationary	7	Numerical	Standardized	OpenStreetMap contributors (2017)
No. of hotels		Within a certain radius (100 m, 200 m, 500 m, 1000 m, 2000 m, 5000 m, 6000 m)	Stationary	7	Numerical	Standardized	OpenStreetMap contributors (2017)
No. of hospitals		Within a certain radius (100 m, 200 m, 500 m, 1000 m, 2000 m, 5000 m, 6000 m)	Stationary	7	Numerical	Standardized	OpenStreetMap contributors (2017)
No. of industries		Within a certain radius (100 m, 200 m, 500 m, 1000 m, 2000 m, 5000 m, 6000 m)	Stationary	7	Numerical	Standardized	OpenStreetMap contributors (2017)

Continued

Table 6. Continued

Feature name	Further explanation	Spatial scope	Timing for daily model	# Features	Type	Scaling	Data source
Percent of area used for farming		In the planning area	Stationary	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023)
Percent of area used for horticulture		In the planning area	Stationary	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023)
Percent of area used for cemeteries		In the planning area	Stationary	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023)
Percent of area used for waterways		In the planning area	Stationary	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023)
Percent of area used for industry		In the planning area	Stationary	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023)
Percent of area used for private gardening		In the planning area	Stationary	1	Numerical	Standardized	Senate Department for Urban Development,

Continued

Table 6. *Continued*

Feature name	Further explanation	Spatial scope	Timing for daily model	# Features	Type	Scaling	Data source
Percent of area used for parks		In the planning area	Stationary	1	Numerical	Standardized	Building and Housing (2023) Senate Department for Urban Development, Building and Housing (2023)
Percent of area used for traffic areas		In the planning area	Stationary	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023)
Percent of area used for forests		In the planning area	Stationary	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023)
Percent of area used for residential housing		In the planning area	Stationary	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023)
Cocioeconomic indicators							
Population density	Inhabitants/km ²	In the planning area	Yearly	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023), Berlin-Brandenburg

Continued

Table 6. *Continued*

Feature name	Further explanation	Spatial scope	Timing for daily model	# Features	Type	Scaling	Data source
Number of inhabitants		In the planning area	Yearly	1	Numerical	Standardized	Office of Statistics (2023), Senate Department for Urban Development, Building and Housing (2023) and Berlin-Brandenburg Office of Statistics (2023)
Average age		In the planning area	Yearly	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023) and Berlin-Brandenburg Office of Statistics (2023)

Continued

Table 6. *Continued*

Feature name	Further explanation	Spatial scope	Timing for daily model	# Features	Type	Scaling	Data source
Gender distribution		In the planning area	Yearly	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023) and Berlin-Brandenburg Office of Statistics (2023)
Share of population with migration background		In the planning area	Yearly	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023) and Berlin-Brandenburg Office of Statistics (2023)
Share of foreigners (total, EU-foreigners, non-EU-foreigners)		In the planning area	Yearly	3	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023) and Berlin-Brandenburg Office of Statistics (2023)

Continued

Table 6. Continued

Feature name	Further explanation	Spatial scope	Timing for daily model	# Features	Type	Scaling	Data source
Share of population unemployed		In the planning area	Yearly	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023) and Berlin-Brandenburg Office of Statistics (2023)
Share of population with tenure exceeding 5 years		In the planning area	Yearly	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023) and Berlin-Brandenburg Office of Statistics (2023)
Net migration rate (moving to/away from the area)		In the planning area	Yearly	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023) and Berlin-Brandenburg Office of Statistics (2023)
Age-specific demographic		In the planning area	Yearly	2	Numerical	Standardized	Senate Department for Urban

Continued

Table 6. *Continued*

Feature name	Further explanation	Spatial scope	Timing for daily model	# Features	Type	Scaling	Data source
proportions (individuals aged <18 & ≥ 65)							Development, Building and Housing (2023) and Berlin-Brandenburg Office of Statistics (2023)
Greying index		In the planning area	Yearly	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023) and Berlin-Brandenburg Office of Statistics (2023)
Birth rate		In the planning area	Yearly	1	Numerical	Standardized	Senate Department for Urban Development, Building and Housing (2023) and Berlin-Brandenburg Office of Statistics (2023)
Weather							
Average temperature	In °C	Stationary	Daily	1	Numerical	Standardized	meteostat (2022)
Daily maximum temperature	In °C	Stationary	Daily	1	Numerical	Standardized	meteostat (2022)

Continued

Table 6. Continued

Feature name	Further explanation	Spatial scope	Timing for daily model	# Features	Type	Scaling	Data source
Daily minimum temperature	In °C	Stationary	Daily	1	Numerical	Standardized	meteostat (2022)
Precipitation	In mm	Stationary	Daily	1	Numerical	Standardized	meteostat (2022)
Maximum snow depth	In mm	Stationary	Daily	1	Numerical	Standardized	meteostat (2022)
Sunshine duration	In minutes	Stationary	Daily	1	Numerical	Standardized	meteostat (2022)
Average wind speed	In km/h	Stationary	Daily	1	Numerical	Standardized	meteostat (2022)
Wind direction	In degrees	Stationary	Daily	1	Numerical	Standardized	meteostat (2022)
Peak wind gust	In km/h	Stationary	Daily	1	Numerical	Standardized	meteostat (2022)
Average sea-level air pressure	In hPa	Stationary	Daily	1	Numerical	Standardized	meteostat (2022)
Motorized traffic							
No. of vehicles/cars/lorries		Within a 6 km radius to the counter	Daily	3	Numerical	Standardized	Berlin Open Data (2022b)
No. of vehicles/cars/lorries		Within the whole city	Daily	3	Numerical	Standardized	Berlin Open Data (2022b)
Speed of vehicles/cars/lorries		Within a 6 km radius to the counter	Daily	3	Numerical	Standardized	Berlin Open Data (2022b)
Speed of vehicles/cars/lorries		Within the whole city	Daily	3	Numerical	Standardized	Berlin Open Data (2022b)

A.5. Overview counting stations

This section provides further details on the counting stations, including an overview of the available counting stations (Table 7) and a mapping of them (Figure 5), as well as some descriptions of the measurements (Figure 6).

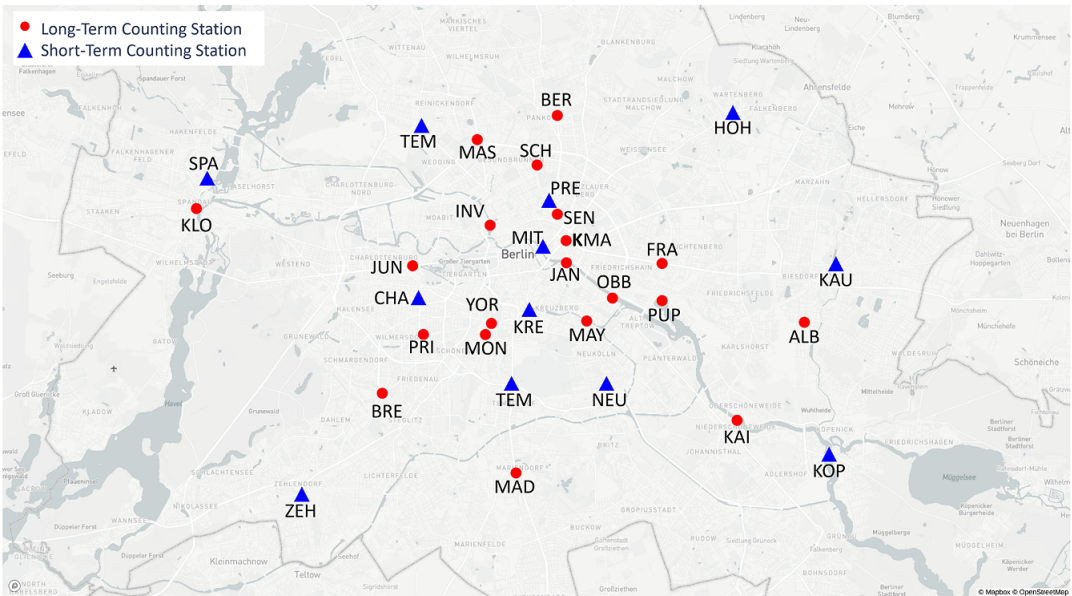


Figure 5. Location of the 12 short-term and 20 long-term counting stations within Berlin.

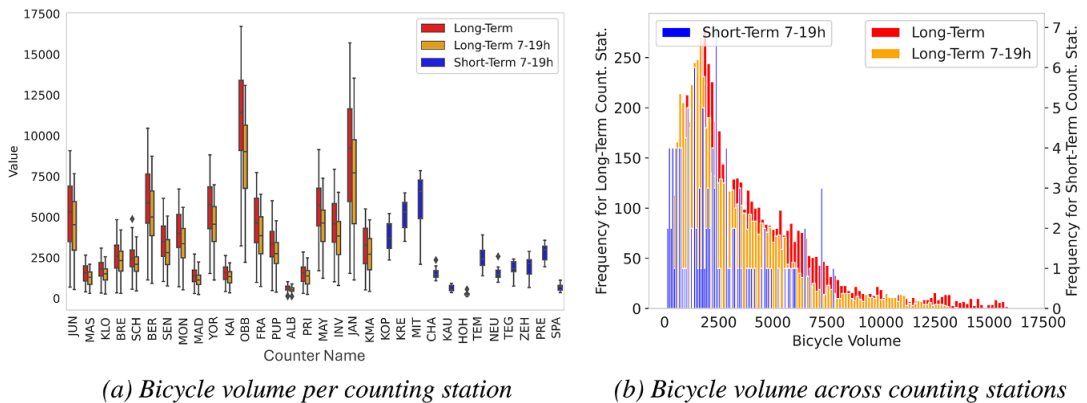


Figure 6. Descriptive statistics of the daily counter measurements (number of bicycles per day).

The boxplots and histograms of counting stations' measurements reveal distinct patterns. The boxplot (Figure 6a) demonstrates that long-term stations have very infrequent high values. Conversely, short-term stations show a lower mean count with few high values. This disparity arises as they cover a shorter period, including fewer days with extreme events. Short-term stations consider only daytime measurements (7–19 h), omitting the lower nighttime counts. This assumption is supported by Figure 6b, depicting permanently higher values for the long-term, in comparison to the long-term 7–19 h. The distributions, exhibit a right-skewed, long-tailed pattern, occasionally indicating notably high cycling volumes. However, the right-skewedness is less pronounced for short-term stations.

Table 7. *The long-term and short-term counting stations*

Counter name ¹	Location	Two-way combined to one-way ²	Installed in	# 0 h–24 h measurements ³	# 7 h–19 h measurements ³
Long-term counting stations (20 locations)					
JAN	Jannowitzbrücke	True	2015	380	381
BRE	Breitenbachplatz	True	2016	380	381
PRI	Prinzregentenstraße	False	2015	380	381
FRA	Frankfurter Allee	True	2016	381	382
BER	Berliner Straße	True	2016	379	380
SCH	Schwedter Steg	False	2012	376	378
MON	Monumentenstraße	False	2015	373	376
MAY	Maybachufer	False	2016	380	381
KAI	Kaisersteg	False	2016	379	380
MAS	Markstraße	False	2015	376	377
MAD	Mariendorfer Damm	True	2016	374	376
KLO	Klosterstraße	True	2016	373	374
PUP	Paul-und-Paula-Uferweg	False	2015	351	352
ALB	Alberichstraße	False	2015	355	356
OBB	Oberbaumbrücke	True	2015	218	218
INV	Invalidenstraße	True	2015	204	205
YOR	Yorckstraße	True	2015	179	180
KMA	Karl-Marx-Straße	False	2021	164	164
JUN	Straße des 17. Juni	True	2021	164	164
SEN	Senefelderplatz	False	2022	53	53
Short-term counting stations (12 locations)					
CHA	Joachimsthaler Str./ Lietzenburger Str.	True	2001	0	10
TEM	Tempelhofer Damm	False	2011	0	10
TEG	Scharnweberstraße	False	2011	0	10
SPA	Schönwalder Str. Neuendorfer Str.	True	2001	0	10
KRE	Zossener Str./ Blücherstr.	True	2001	0	10
HOH	Pablo-Picasso-Str./ Falkenseer Chaussee	True	2001	0	10
KOP	Lange Brücke	True	2001	0	10
PRE	Kastanienallee/Schwedter Str.	True	2001	0	10
NEU	Karl-Marx-Straße	False	2011	0	10
MIT	Karl-Liebnecht-Str./ Spandauer Str.	True	2001	0	10
ZEH	Teltower Damm/Schönower Straße	True	2001	0	10
KAU	Altentrepptower Straße	False	2011	0	9

¹We use the abbreviation throughout the paper, to pinpoint the individual counters.²In some locations the counter stations count the passing bikes independently for the different sides of the street. In these cases, we combined them, ignoring the direction of the flow.³The number of observations refers to the time span in which all other necessary features were available.

A.6. Infrastructure features: city-wide in comparison to counting station locations

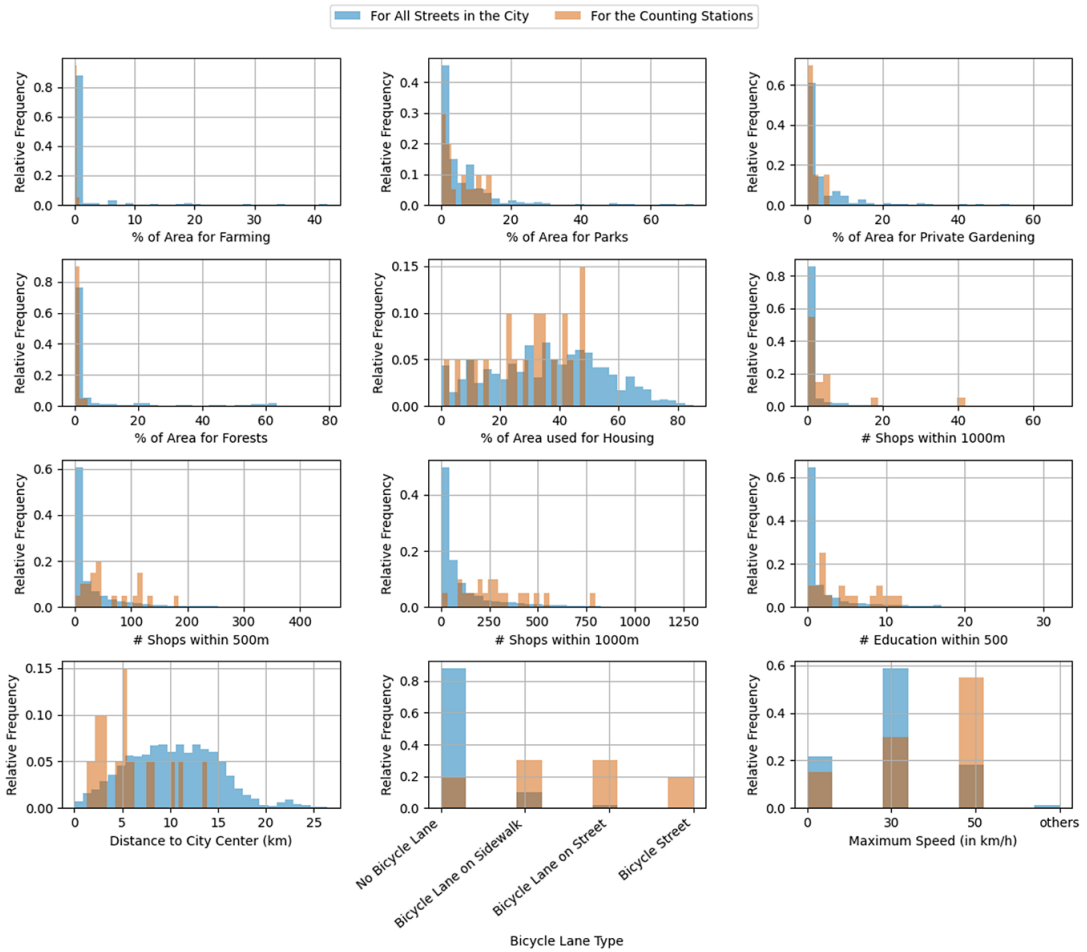


Figure 7. Frequency histograms of selected infrastructure features, illustrating the distribution of these features for both counting station locations and all street segments across Berlin. A street segment describes a street section between two intersections/the end of a street and an intersection. It should be noted that since paths in forests and parks are included in the analysis, there is a higher-than-expected proportion of segments with a speed limit of 0 km/h.

A.7. Feature engineering bike-sharing data

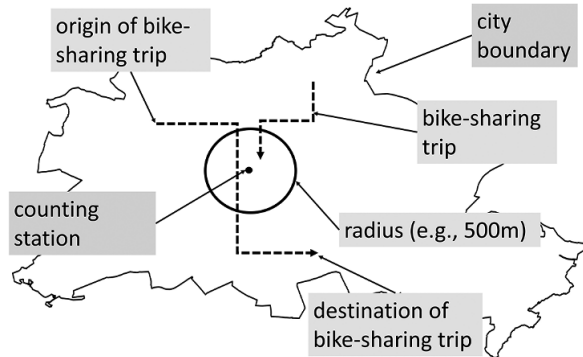
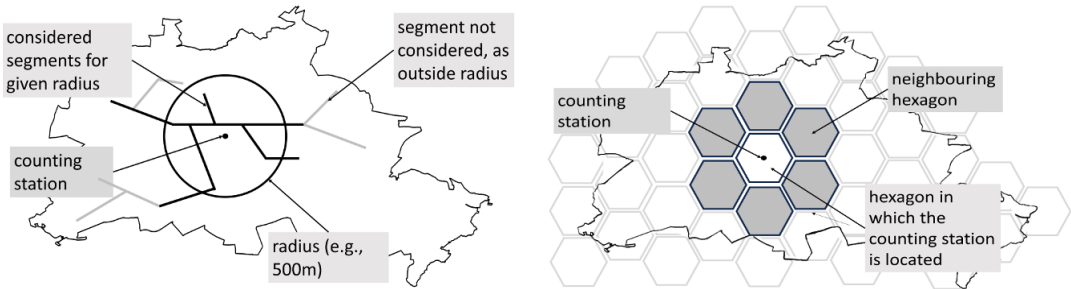


Figure 8. The bike-sharing data was feature engineered based on a radius: For a given day all passing bike-sharing trips passing, starting or ending within a certain radius around the counting station in question were counted. This was also done for the entirety of the city. In this visualization, two bike-sharing trips are depicted. Given that both trips started and ended in a given day, the graph would produce a count of two passing, newly rented, and returned bike trips in the whole city, as well as two passing bike trips in the radius and one ending and zero originating.

A.8. Feature engineering crowdsourced data



(a) All segments lying, partially or fully, within a certain radius of the counting station (black) were considered for the feature engineering. Other segments were not considered (grey). (b) The hexagon in which the counter is located (white, in the center) and its neighboring entities (grey).

Figure 9. The Strava data, both the hexagon and the street segment data, was feature-engineered. We computed the average across features for both data types, considering observations within a certain proximity. For the street segment data, we considered all segments within a certain radius. For the hexagon, we took the average of the features across the six neighboring hexagons. Additionally, we included the features for the hexagon, where the relevant counting station is located.

A.9. Comparison crowdsourced and bike-sharing data

Table 8. Descriptives of the Strava and bike-sharing data. All specifications are in percent of the total trips recorded. The numbers indicate that bike-sharing trips are more evenly conducted throughout the whole day. Also, bike-sharing riders are much slower on average than Strava users, which seems reasonable, given the different quality of bike-sharing trips versus private bikes

Group	Type	Strava	Bike-sharing
Type of bike	Non e-bike trips	99.83%	100%
	E-bike trips	0.17%	0%
Purpose of trip ¹	Commute trips	39.74%	NA
	Leisure trips	60.26%	NA
Sex of user	Male	89.57%	NA
	Female	9.69%	NA
	Gender unspecified	0.74%	NA
Age of user	18–34	33.63%	NA
	35–54	62.47%	NA
	55–65	3.67%	NA
	65+	0.23%	NA
Time of trip ²	Morning	24.82%	23.65%
	Midday	27.88%	27.09%
	Evening	40.64%	32.88%
	Night	6.66%	16.38%
Basic parameters	Average distance	NA	2.90 km
	Average duration	NA	24.31 min
	Average speed	21.14 km/h	11.05 km/h

¹Is categorized by Strava. Strava identifies commutes through a model This model utilizes the “commute” tag provided by Strava members as ground truth. The term “commuting” encompasses all trips that are not related to leisure activities (Strava Metro, 2023).

²Morning: 5 h–10 h, midday: 10 h–15 h, evening: 15 h–20 h, overnight: 20 h–5 h. The Strava data was categorized by Strava. For the bike-sharing data, we considered the moment of departure.

Usage patterns differ between Strava and bike-sharing. On average bike-sharing usage is more evenly distributed throughout the day, whereas Strava trips are more likely to be recorded during the midday and evening. Also, bike-sharers ride on average with a speed of 11.05 km/h whereas Strava are almost at double the speed with 21.14 km/h. This seems reasonable, as Strava is used heavily to track sporting activities and bike-sharing bikes tend to be of lower quality.

A.10. Strava features: city-wide in comparison to counting station locations

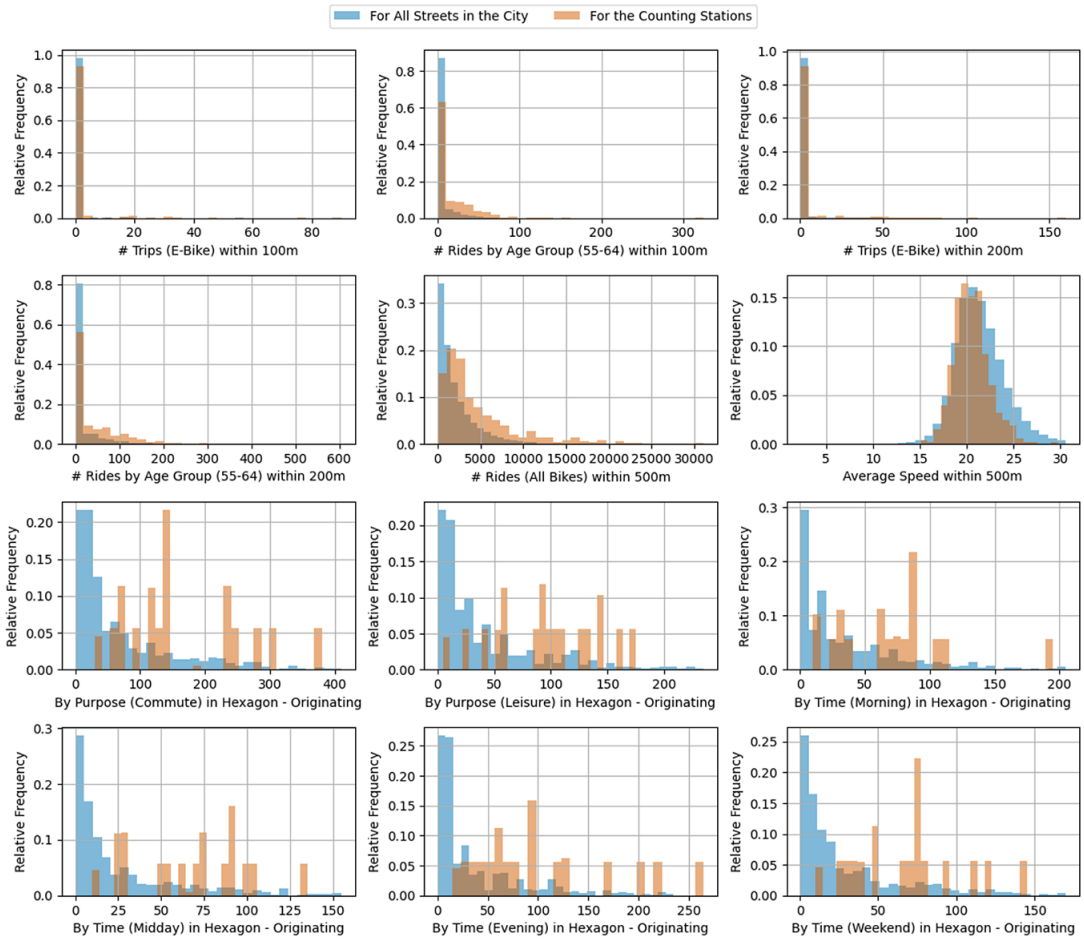


Figure 10. Histograms displaying the frequency distribution of selected Strava features, comparing counting station locations with all street segments across Berlin, exemplified for September 2022. A street segment describes a section of a street that lies between two intersections/the end of a street and an intersection. Some features include high outliers. To enhance readability, we capped values at the 99th percentile.

A.11. Connectivity measures

The degree specifies the number of edges connected to a node v .

$$\text{degree}(v) = \text{number of edges incident to } v$$

Betweenness centrality of a node v , indicates the extent to which it lies on the shortest path between pairs of other nodes s and t . With σ_{st} the total number of shortest paths from s to t and $\sigma_{st}(v)$ the number of those paths passing via v .

$$\text{betweenness}(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Closeness of a node, indicates how quickly all other n nodes can be reached in the graph, with $d(u, v)$ the shortest path distance between u and v .

$$\text{closeness}(v) = \frac{1}{n-1 \sum_{u \neq v} d(v, u)}$$

The clustering coefficient measures the degree to which nodes in the neighborhood of node v are connected, providing insight into the local density. With the number of triangles centered at v referring to the closed loops of length 3 that include node v and its neighbors:

$$C(v) = \frac{2 \times \text{number of triangles centered at } v}{\text{degree}(v) \times (\text{degree}(v) - 1)}$$

A.12. Location of motorized traffic counting stations

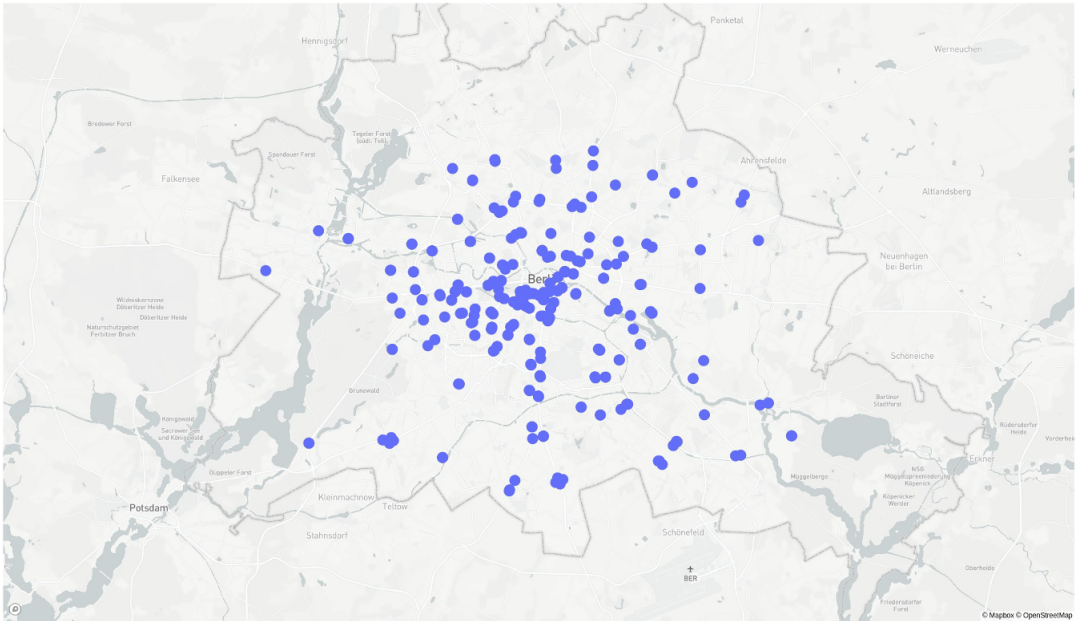


Figure 11. Location of counting stations measuring motorized traffic.

A.13. QQ plot of the target feature before and after log-transformation

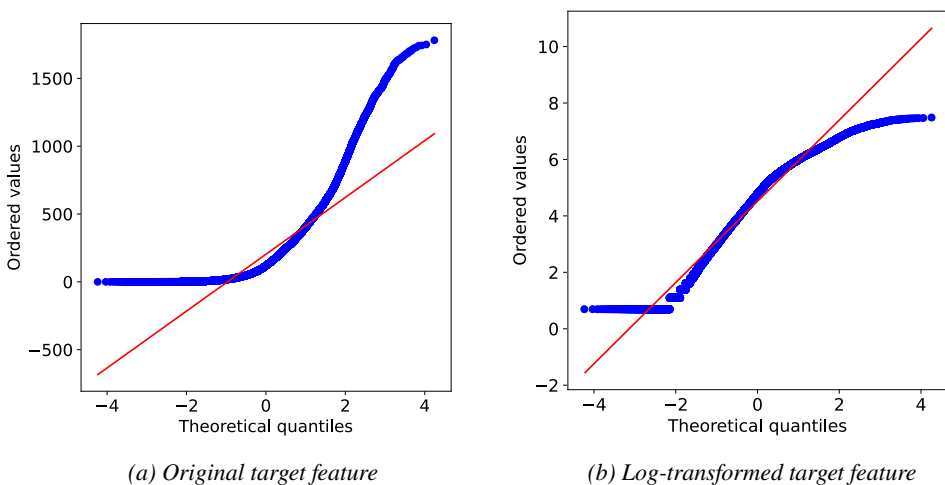


Figure 12. QQ Plot of the target feature (measurements of the counting stations), before and after the log-transformation.

A.14. ML models' hyperparameters**Table 9.** *Models and their tuned hyperparameters*

Models	Tuned hyperparameters
Linear regression	–
Decision tree	Maximum depth, minimum samples for splitting, min samples in leaf Node, splitting criterion
Gradient boosting	Number of estimators, learning rate, maximum depth, minimum samples for splitting, minimum samples in leaf node
XGBoost	Learning rate, maximum depth, subsample size, column subsampling rate, minimum child weight, gamma
Random forest	Number of estimators, maximum depth, minimum samples for splitting, bootstrap sampling
Support vector regression	C (regularization parameter), kernel choice, degree (for polynomial kernel), gamma, epsilon
Shallow neural network	Hidden layer sizes, activation function, learning rate

A.15. Feature selection methods**Table 10.** *Models, the applied feature selection method, and the number of selected features for both the all day and the 7-19 h specification*

Model	Feature election method	# Features selected all day	# Features selected 7-19 h
Linear regression	Recursive feature elimination with cross-validation and linear regression	9	9
Decision tree	Sequential feature selection with decision tree	38	38
Gradient boosting	Recursive feature elimination with cross-validation and gradient boosting	195	105
XGBoost	Sequential feature selection with XGBoost	38	38
Random forest	Recursive feature elimination with cross-validation and random forest	15	215
Support vector regression	Select K Best	28	28
Shallow neural network	Select K Best	28	28

Of the 257 available features (see Table 2), 243 are selected at least once across any feature selection approaches we applied. The selection frequency varies significantly, with some features chosen for many models while most are selected only once or twice. The individual most frequently chosen features are the population density, the age-specific demographic proportions (>65 years), and the count of shops within 2000 m. Other relatively frequently selected characteristics include the categories of infrastructure

characteristics (points of interest, such as the number of stores and the use of the area, especially the proportion of the area used for parks and agriculture), Strava characteristics (traffic flow in the area for small and medium radii) and motorized traffic characteristics (city-wide number of cars). For each model, features based on smaller radii tend to be selected more frequently, but features based on larger radii are also included for all models.

Cite this article: Kaiser SK, Klein N and Kaack LH (2025). From counting stations to city-wide estimates: data-driven bicycle volume extrapolation. *Environmental Data Science*, 4: e13. doi:10.1017/eds.2025.5