From Counting Stations to City-Wide Estimates: Bicycle Volume Extrapolation With Multi-Source and Sample Data

Silke K. Kaiser*[1], Nadja Klein[2], and Lynn H. Kaack[3]

[1]PhD Student, Hertie School Berlin, Germany
[2]Full Professor, Technische Universität Dortmund, Germany
[3]Assistant Professor, Hertie School Berlin, Germany

Short summary

Switching to cycling in urban areas reduces greenhouse gas emissions and improves the health of society as a whole. In order to promote cycling as a mode of transport, accurate information on the volume of passing bicycles is essential for cities to plan infrastructure development strategically. Currently, most cities can only rely on data from sparsely located counting stations. To address this problem, we extrapolate data from these stations to estimate city-wide bicycle volumes for Berlin. Our work involves machine learning models and various public data sources, including app-based crowdsourcing, bike sharing, motorized traffic data, and more. In addition, we simulate performance improvements by conducting sample counts at predicted locations. By providing the model with ten days of count samples for the predicted locations, we can cut the error in half and significantly minimize the variation in performance between predicted locations.
Keywords: City-wide flow, big data analytics, bicycle flow, machine learning, cycling design.

1 Introduction

Cycling offers several health benefits, such as improved cardiorespiratory health and reduced risk of cancer mortality, while also contributing to public health by reducing emissions and improving air quality (Oja et al., 2011; Woodcock et al., 2009). Additionally, shifting from motorized transport to bicycles and e-bikes helps mitigate climate change by reducing greenhouse gas emissions (H.-O. Pörtner, D.C. Roberts, E.S. Poloczanska, K. Mintenbeck, M. Tignor, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem (eds.), 2022). Infrastructure improvements play a key role in promoting urban cycling: Not only do separated bike lanes improve safety (Morrison et al., 2019), but people's perceived risks align with actual risks, especially among adult cyclists and women, who prefer designated bike lanes (Dill, 2009; Garrard et al., 2008). Contributing to the shift to non-motorized transport requires targeted infrastructure improvements (Olmos et al., 2020; Larsen et al., 2013). However, given scarce financial resources and limited public space, data-driven approaches are essential to inform policymakers and encourage goal-oriented changes.

A critical aspect of data-driven strategies is the availability of bicycle volume data. Currently, such data is collected by sparse and costly counting stations (Ryus et al., 2014). This study aims to extrapolate this data to citywide street-level estimates by combining machine learning (ML) methods with various publicly available data sources and sample counts.

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 that have been available for some time, such as weather (Miranda-Moreno & Nosal, 2011), infrastructure (Strauss & Miranda-Moreno, 2013), socioeconomic information (Miah et al., 2022), and vacation data (Holmgren et al., 2017). With the recent widespread adoption of smartphones, new data sources have emerged, such as crowdsourced information from the Strava application (Lee & Sener, 2021) and bike-sharing logs (Miah et al., 2022). Additionally, previous studies have explored the extrapolation of bicycle volumes using only some data sources and classical regression approaches (Roy et al., 2019; Sanders et al., 2017; Dadashova & Griffin, 2020). Likewise, ML has gained increasing interest in the last decade for extrapolating motorized traffic (Sekuła et al., 2018; Das & Tsapakis, 2020; Zahedian et al., 2020). However, to our knowledge, studies have yet to combine ML methods with as many different data sources to provide reliable, fine-grained predictions of bicycle counts beyond available counting stations.

To address this research gap, this paper focuses on predicting bicycle volumes at unseen locations using ML models and different data sources. The study is conducted in Berlin, Germany, with a bicycle modal share of 42%, which is close to the European average of 37% (European Metropolitan Transport Authorities, 2021), making it a suitable case. We answer two key questions: First, can bicycle volumes be predicted at unseen locations using different data sources? Second, how does performance improve with additional sample counts for predicted locations?

2 Methodology

Data

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 more locations where short-term counts are conducted on several days throughout the year (Senate Department for the Environment, Mobility, Consumer and Climate Protection Berlin, 2022). To accurately predict bicycle counts, we use information from several additional data sources. These include data on infrastructure, socioeconomic factors, motorized traffic, weather, holidays, bike sharing, and data from a cyclist tracking application (Strava application). Bike sharing and Strava data directly capture bicycle traffic but serve different user bases. The former provides details on the precise timing and origin-destination pairs of individual trips taken on short-term and dockless rental bikes. At the same time, the latter consists of anonymized georeferenced data aggregated to provide trip counts for regions and road segments between intersections by tracking user movements. Bike-sharing, crowdsourced, and motorized traffic data are feature-engineered to provide counts of passing bikes and cars within different radii around a counter and the day in question. Socioeconomic and infrastructure features are assigned based on the location of counting stations. A comprehensive list of all features is available in the table 1, along with references to the respective data sources. Given the abundance of features, we implement model-specific feature selection (FS) to reduce computational requirements and potentially improve model performance. Each model is tested using univariate FS with SelectKBest, recursive feature elimination based on XGBoost, SelectFromModel with XGBoost, and FS via sequential selection with linear regression (Pedregosa et al., 2011). As the bike-sharing data covers only the periods from April to December 2019 and June to December 2022, we limit our study to these periods, largely excluding the timeframe of the COVID-19 pandemic and its impact on transportation.

Table 1: The table gives a summary of the features per data source used in this study.

Data	Description of features	No of features	Data source
Time	Month, day of month, weekday, weekend, year	5	inherent
Holiday	School holiday, public holiday	2	Senate Department for Education, Youth and Family (n.d.)
Bike-sharing	Number of bicycles originated, returned rented within various radii*	18	CityLab Berlin (n.d.) and Nextbike (2020)
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 *	91	Strava Metro (n.d.)
Infrastructure	Latitude, longitude, distance to city center, maximum speed, bicycle lane type, number of shops/education centers/hotels/hospitals for various radii,* percent of area used for farming/horticulture/cemeteries/waterways/industry/ private gardening/parks/traffic areas/forests/ residential housing	31	OpenStreetMap contributors (2017); Senate Department for Urban Development, Building and Housing (n.d.)
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 (n.d.); Berlin-Brandenburg Office of Statistics (2020)
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 (n.d.)
Motorized traffic	total number and speed of vehicles/cars/lorries within different radii**	12	Berlin Open Data (2022)
Total no of features		184	

* The features are each computed for a radius of 0.5, 1, 2, and 5km.
** The features are each computed for a radius of 6km.

Methodology

First, to answer the question of how well bicycle volume can be predicted at unseen locations using different data sources, we employ the following strategy. We compare several classical machine learning algorithms, including Linear Regression, Decision Tree, Random Forests, Gradient Boost, XGBoost, Support Vector Machine, and a Shallow Neural Network. For detailed information on the models, we refer the reader to Géron (2022). We tune their hyperparameters with random search. We evaluate the predictions on a daily and annual scale. 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 variations, 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. Average Annual Daily Bicycle Volume (AADB) is the average number of bicycles passing a given location per day for a given year. We calculate performance for the AADB by predicting daily counts and evaluating their average against the annual ground truth average. To simulate extrapolation, we evaluate our models using leave-one-group-out (LOGO) cross-validation. The method follows the same principle as standard cross-validation 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 counting station, and then evaluated on the hold-out count station. In addition, we use each short-term station as test data for a model trained on all long-term stations. We provide the average error across stations, which implies that each location in the test data is equally weighted. When computing these predictions, it is important to note that the hourly long-term data are measured from 0h-24h, while the short-term counts are only performed from 7h-19h. To ensure compatibility, we train the model predicting the short-term stations only on daily measurements, which are computed as the sum of the 7h-19h hourly measurements. We also perform the analysis of the long-term stations on daily measurements based on 0h-24h and 07-19h data separately. The former allows us to infer diurnal effects for long-term stations, and the latter can be used to compare results with short-term counting stations. To provide information on the absolute and relative size of our errors. We pursue two strategies. We include a baseline computed using the mean of the training data as a prediction. Also, we use the symmetric mean absolute percentage error (SMAPE) as an evaluation metric. SMAPE is defined as follows, where n is the number of observations, y_i is the true value, and $\hat{y}_i$ is the prediction of the variable of interest:

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

Second, we use the following methodology to evaluate how performance improves with additional sample counts for predicted locations: Since research has shown that at least a 24-hour window should be used when collecting sample data (Nordback et al., 2013), and suggests that even longer periods are better for estimating annual volumes (Hankey et al., 2014; Nosal et al., 2014; Nordback et al., 2013), we choose to simulate three different sample data collection strategies. In the first strategy, data collection is commissioned for each location for one day at a time (1-day). The days are chosen randomly throughout the year. In the second and third strategies, we simulate data collection for three (3-day) or seven (7-day) consecutive days. These multi-day periods are randomly distributed throughout the year. In total, we simulate collecting data for up to 28 days. We evaluate the collection of this sample data by iterating over the count stations. Each counting station serves once as a new ("hold-out") location. For that location, we set aside some of the available data to represent sample counts taken at that location according to the

sampling strategies described above (1, 3, or 7 days). We use the remaining data from that location as the test set. For training, we implement two modeling scenarios. First, we train the model on both the sample data and the data from the other counting stations. We give a weight of 25% to the sample counts and 75% to the long-term station data. This scenario benefits from both sample and city-wide long-term information. Therefore, we call it the "full-city" scenario. Second, we train the model using only the sample data. Since it only uses information from the location in question, we call this model the "location-specific" scenario. We then use both models to perform prediction on the test set. We repeat this process for each counting station and compute the average of the resulting errors. We train and evaluate the models after each additional day of data collection. This allows for a continuous comparison of the different approaches over time, e.g. after the second day of data collection. Finally, we repeat this procedure 10 times with different sample days to allow for uncertainty estimation and provide 95% confidence intervals. In addition, we include a baseline that shows the error in predicting the site-specific volume as the mean of the sample data collected at that site. In this way, we want to show that the inclusion of multi-source data is still relevant when obtaining sample counts. It is important to note that we simulate the collection using only 19 long-term counting stations, as all short-term and one long-term station have too few observations per site. As before, we use the XGBoost model with SMAPE.

3 Results and discussion

Table 2: SMAPE error for the various machine learning models at the daily, and average annual daily bicycle volume (AADB) level. The gray background implicates the columns employed as the criterion for model selection.

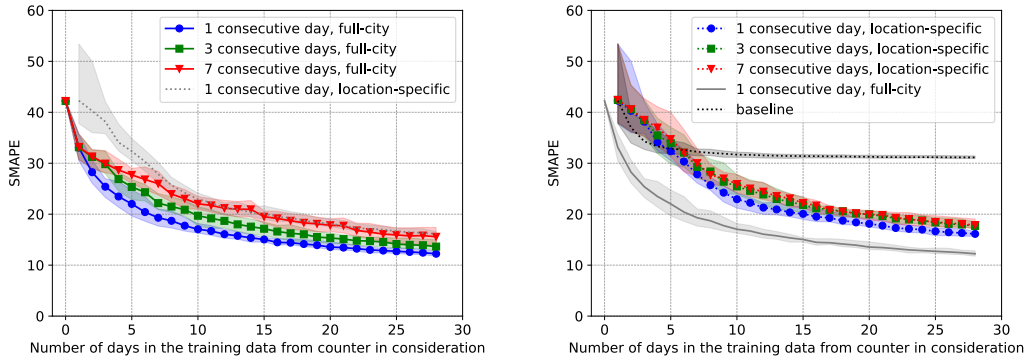
Dimension	daily	daily	daily	AADB	AADB	AADB
Time	0h-24h	7h-19h	7h-19h	0h-24h	7h-19h	7h-19h
Counter type	long-term	long-term	short-term	long-term	long-term	short-term
Evaluation	LOGO	LOGO	test	LOGO	LOGO	test
	(1)	(2)	(3)	(4)	(5)	(6)
Linear regression	127.92	117.68	95.80	126.83	114.11	86.79
Decision tree	51.72	51.85	47.50	44.51	47.10	**43.73**
Random forest	46.04	46.47	**46.97**	41.63	**42.88**	48.02
Gradient boosting	43.21	47.14	64.81	40.01	44.08	59.04
XGBoost	**41.24**	**46.67**	70.64	**38.86**	44.27	67.38
Support vector machine	47.06	48.12	55.38	44.85	43.88	47.70
Shallow neural network	57.22	56.99	70.10	52.86	53.22	63.09
Baseline	66.62	65.67	70.62	66.62	65.67	70.62

Regarding the ability to extrapolate bicycle volume at unseen locations using multi-source data, we find the following: We observe that ensemble methods (XGBoost, gradient boosting, random forest, and decision trees) outperform the baseline, support vector machines, linear regression, and shallow neural networks (Table 2). We identify XGBoost as the top-performing model through LOGO analysis involving all long-term counting stations and the 0h-24h data. It's important to note that although this model doesn't yield the lowest errors in short-term count predictions, we aim to demonstrate its extrapolation capabilities for bicycle volume across any street through a proof of concept. The animated version of this proof of concept is accessible

online at https://silkekaiser.github.io/research.

For a more detailed evaluation of the XGBoost model's performance, we examine the SMAPE variation between stations. On a daily basis, the model exhibits strong performance for over half of the stations (with a SMAPE around 20). However, for some stations, the SMAPE exceeds 80, and there is significant variability in performance across counters in the AADB. The poorly performing sites consistently display high variance in measurements, with each site being consistently either overpredicted or underpredicted. Despite our comprehensive inclusion of various features from existing literature, our analysis fails to uncover common characteristics among the worst-performing counters that would pinpoint the model's failures. Consequently, we conclude that latent factors within the data generation process remain unaccounted for. To address this issue, we plan to explore potential mitigations using different sample sizes.

Furthermore, we want to elaborate on our results regarding the performance improvement of sample data collection: in all scenarios, as well as for both the full-city and the location-specific scenarios, sample data collection significantly improves the prediction performance for new locations (Figure 1). Among the different collection strategies, using a finer granularity (1-day) turns out to be most superior. This superiority is more pronounced for the full-city than for the location-specific scenario. The advantage of collecting data on as many different days as possible is further emphasized by the fact that there are discernible error decreases after the 7th and 14th days as well as after the 3rd and 6th days for the 7-day and 3-day strategies (1a). Furthermore, our analysis reveals a consistent superiority of the full-city scenario employing the 1-day strategy over the location-specific scenario. A comparison of the 1-day strategy between these scenarios



(a) Combined sample and long-term data for training (full-city)

(b) Sample data only for training (location-specific)

Figure 1: Shown is 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.

indicates that, to achieve a SMAPE of 20, one needs an average of 7 days of sample data in the full-city, as opposed to 14 days in the location-specific scenario. This highlights the substantial advantages models can gain from information gathered at locations beyond the specific one in focus. Additionally, we observe the significant relevance of utilizing multi-source data when working with sample data. The baseline SMAPE never drops below 30 (as depicted in Figure 1b), whereas the full-city appraoch attains an average error of around 17 after 10 days of sample counts. This underscores the crucial role of incorporating multi-source data alongside sample counts in improving model performance.

For further comparison with the above results, we train an XGBoost model on the full-city scenario in combination with multi-source data and ten days of sample counts using the 1-day strategy. We collect these ten days randomly across all observations and across both years. Again, to account for the randomness of the sample data collection, we calculate 10 replicate samples and take the average of these. We find that we are able to predict new locations at the daily level with an average SMAPE of 17.44 and MAE of 594.59. For the AADB we get 11.94 and 360.84 respectively. On closer inspection, we also find that these errors vary little between the stations. 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, but also more reliable.

4 Conclusions

In our study, we use multi-source data to predict bicycle volume at unseen locations using XGBoost, achieving a SMAPE of 41.24 for daily estimates and 40.41 for average annual daily bicycle (AADB) estimates. While our approach allows for cost-effective estimation of bicycle volumes for entire cities, the error remains relatively high. Examination reveals considerable variability in error across locations, with half of the locations having a SMAPE about half the average and a few extreme locations having a SMAPE twice the average. We conclude that there are latent factors in the location data generation processes that remain unaccounted for. Therefore, we chose to simulate the incorporation of sample counts at unseen locations. Ten days of sample counts significantly reduces error and variance across locations. Incorporating multi-source data and information from other locations reduces this error to 17.44 for daily estimation and 11.94 for AADB.

However, recognizing the additional cost of sample counts, we suggest that future studies explore more complex modeling approaches that account for spatial and temporal dependencies as well as location- and time-specific effects. Another limitation lies in the spatial homogeneity of available data points within Berlin, suggesting consideration of a more diverse selection of spatial points for enhanced modeling incorporating spatial aspects. Extending our research to several cities could reveal patterns in different urban settings.

In conclusion, our research demonstrates the feasibility of estimating citywide bicycle volumes by integrating open-source data with permanent and sample counts using ML algorithms. This approach offers the potential for accurate street-level estimates to improve cycling conditions and infrastructure, thereby promoting cycling as a sustainable mode of transportation in cities. As cities increasingly prioritize cycling, the importance of estimating citywide bicycle volumes is likely to grow, providing exciting research opportunities, particularly in multi-city studies with rich ground-truth data.

Acknowledgements

We thank E. Kolibacz for his technical support. We are grateful to CityLab Berlin for providing their bike-sharing data. The European Union's Horizon Europe research and innovation program funded this project 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).

References

Berlin-Brandenburg Office of Statistics. (2020). *Kommunalatlas Berlin [Municipal Atlas Berlin].* Retrieved 2023-01-09, from `https://instantatlas.statistik-berlin-brandenburg.de/instantatlas/interaktivekarten/kommunalatlas/atlas.html`

Berlin Open Data. (2022). *Verkehrsdetektion Berlin [Traffic detection Berlin].* Retrieved 2023-03-01, from `https://daten.berlin.de/datensaetze/verkehrsdetektion-berlin`

CityLab Berlin. (n.d.). Shared Mobility Flows. , *Accessed: 01.03.2022*. Retrieved 2022-03-01, from `https://bikesharing.citylab-berlin.org/,https://github.com/technologiestiftung/bike-sharing`

Dadashova, B., & Griffin, G. P. (2020, July). Random parameter models for estimating statewide daily bicycle counts using crowdsourced data. *Transportation Research Part D: Transport and Environment*, *84*, 102368. Retrieved 2023-05-09, from `https://www.sciencedirect.com/science/article/pii/S1361920920305551` doi: 10.1016/j.trd.2020.102368

Das, S., & Tsapakis, I. (2020, March). Interpretable machine learning approach in estimating traffic volume on low-volume roadways. *International Journal of Transportation Science and Technology*, *9*(1), 76–88. Retrieved 2022-10-20, from `https://linkinghub.elsevier.com/retrieve/pii/S2046043019301108` doi: 10.1016/j.ijtst.2019.09.004

Dill, J. (2009, January). Bicycling for Transportation and Health: The Role of Infrastructure. *Journal of Public Health Policy*, *30*(S1), S95–S110. Retrieved 2023-09-27, from `http://link.springer.com/10.1057/jphp.2008.56` doi: 10.1057/jphp.2008.56

European Metropolitan Transport Authorities. (2021). *Barometer 2021 – Based on 2019 data - Summary Version.* Retrieved from `https://www.emta.com/publications/article-emta-barometer-of-public-transport/`

Garrard, J., Rose, G., & Lo, S. K. (2008). Promoting transportation cycling for women: The role of bicycle infrastructure. *Preventive medicine*, *46*(1), 55–59. Retrieved 2023-09-27, from `https://www.sciencedirect.com/science/article/pii/S0091743507003039` (Publisher: Elsevier) doi: 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.".

H.-O. Pörtner, D.C. Roberts, E.S. Poloczanska, K. Mintenbeck, M. Tignor, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem (eds.). (2022). IPCC, 2022: Summary for Policymakers. In *Climate Change 2022: Mitigation of Climate Change. Contribution of Working*

Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [P.R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, J. Malley, (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA.

Hankey, S., Lindsey, G., & Marshall, J. (2014, January). 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. Retrieved 2023-11-02, from http://journals.sagepub.com/doi/10.3141/2468-08 doi: 10.3141/2468-08

Holmgren, J., Aspegren, S., & Dahlströma, J. (2017). Prediction of bicycle counter data using regression. *Procedia Computer Science*, *113*, 502–507. Retrieved 2023-01-06, from https://www.sciencedirect.com/science/article/pii/S1877050917317222 doi: https://doi.org/10.1016/j.procs.2017.08.312

Larsen, J., Patterson, Z., & El-Geneidy, A. (2013, June). 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. Retrieved 2023-09-27, from http://www.tandfonline.com/doi/abs/10.1080/15568318.2011.631098 doi: 10.1080/15568318.2011.631098

Lee, K., & Sener, I. N. (2021). Strava Metro data for bicycle monitoring: a literature review. *Transport reviews*, *41*(1), 27–47. Retrieved from https://www.tandfonline.com/doi/full/10.1080/01441647.2020.1798558?casa_token=shftrQZpRCgAAAAA%3AAPohDEH0J74Ruk4c2AzyoXeRtIXqwgs_A9l7T4KVkwf8PXlTLq8v-31FDRSUSZjS-41qdZH8eo9W (Publisher: Taylor & Francis) doi: https://doi.org/10.1080/01441647.2020.1798558

meteostat. (n.d.). The Weather's Record Keeper. , *Accessed: 01.08.2022*. Retrieved 2022-03-02, from https://meteostat.net/en/

Miah, M. M., Hyun, K. K., Mattingly, S. P., & Khan, H. (2022, April). Estimation of daily bicycle traffic using machine and deep learning techniques. *Transportation*. Retrieved 2022-10-18, from https://link.springer.com/article/10.1007/s11116-022-10290-z

Miranda-Moreno, L. F., & Nosal, T. (2011, January). Weather or Not to Cycle: Temporal Trends and Impact of Weather on Cycling in an Urban Environment. *Transportation Research Record*, *2247*(1), 42–52. Retrieved 2022-02-24, from https://journals.sagepub.com/doi/abs/10.3141/2247-06?casa_token=_TfvkjcmtQgAAAAA:n6NZceCFknvtYU5CgNChRpJakZXFRU8cF9aAQvV3KHQroVRNx1B9CNMG13q_oGUKRtPK1_lGAWG8 doi: https://doi.org/10.3141/2247-06

Morrison, C. N., Thompson, J., Kondo, M. C., & Beck, B. (2019, February). On-road bicycle lane types, roadway characteristics, and risks for bicycle crashes. *Accident Analysis & Prevention*, *123*, 123–131. Retrieved 2022-03-01, from https://linkinghub.elsevier.com/retrieve/pii/S0001457518304238 doi: 10.1016/j.aap.2018.11.017

Nextbike. (2020). *Official Nextbike API Documentation.* Retrieved 2022-06-01, from https://github.com/nextbike/api-doc

Nordback, K., Marshall, W. E., & Janson, B. N. (2013). *Development of estimation methodology for bicycle and pedestrian volumes based on existing counts.* (Tech. Rep.). Colorado. Dept. of

Transportation. Research Branch. Retrieved 2023-11-02, from https://rosap.ntl.bts.gov/view/dot/26644

Nosal, T., Miranda-Moreno, L., Krstulic, Z., & Eng, P. (2014). Incorporating weather: a comparative analysis of Average Annual Daily Bicyclist estimation methods 2. *Transportation Research Record*. Retrieved 2023-11-02, from https://trec.pdx.edu/sites/default/files/Nosal%20et%20al%202014.pdf

Oja, P., Titze, S., Bauman, A., de Geus, B., Krenn, P., Reger-Nash, B., & Kohlberger, T. (2011, August). Health benefits of cycling: a systematic review: Cycling and health. *Scandinavian Journal of Medicine & Science in Sports*, *21*(4), 496–509. Retrieved 2022-01-28, from https://onlinelibrary.wiley.com/doi/10.1111/j.1600-0838.2011.01299.x doi: 10.1111/j.1600-0838.2011.01299.x

Olmos, L. E., Tadeo, M. S., Vlachogiannis, D., Alhasoun, F., Espinet Alegre, X., Ochoa, C., ... González, M. C. (2020, June). A data science framework for planning the growth of bicycle infrastructures. *Transportation Research Part C: Emerging Technologies*, *115*, 102640. Retrieved 2023-09-27, from https://www.sciencedirect.com/science/article/pii/S0968090X19306436 doi: 10.1016/j.trc.2020.102640

OpenStreetMap contributors. (2017). *Planet dump retrieved from https://planet.osm.org*. Retrieved from https://www.openstreetmap.org

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, ... E. Duchesnay (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research}*,, *12*, 2825–2830.

Roy, A., Nelson, T. A., Fotheringham, A. S., & Winters, M. (2019). Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists. *Urban Science*, *3*(2), 62. Retrieved from https://www.mdpi.com/2413-8851/3/2/62 (Publisher: MDPI) doi: https://doi.org/10.3390/urbansci3020062

Ryus, P., Ferguson, E., Laustsen, K. M., Schneider, R. J., Proulx, F. R., Hull, T., ... National Academies of Sciences, Engineering, and Medicine (2014). *Guidebook on Pedestrian and Bicycle Volume Data Collection*. Washington, D.C.: Transportation Research Board. Retrieved 2023-01-09, from https://www.trb.org/Publications/Blurbs/171973.aspx

Sanders, R. L., Frackelton, A., Gardner, S., Schneider, R., & Hintze, M. (2017, January). 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. Retrieved 2023-08-30, from http://journals.sagepub.com/doi/10.3141/2605-03 doi: 10.3141/2605-03

Sekuła, P., Marković, N., Laan, Z. V., & Sadabadi, K. F. (2018, October). Estimating Historical Hourly Traffic Volumes via Machine Learning and Vehicle Probe Data: A Maryland Case Study. *arXiv:1711.00721 [cs, stat]*. Retrieved 2022-02-25, from http://arxiv.org/abs/1711.00721 (arXiv: 1711.00721)

Senate Department for Education, Youth and Family. (n.d.). Ferientermine [Vacation Dates]. , *Accessed: 02.08.2022*. Retrieved from https://www.berlin.de/sen/bjf/service/kalender/ferien/artikel.420979.php

Senate Department for the Environment, Mobility, Consumer and Climate Protection Berlin. (2022, December). Jahresdatei geprüfter Rohdaten der Radzählstellen [Annual file of audited raw data of bicycle counting stations]. , *Accessed: 01.03.2023*. Retrieved 2023-03-01, from https://www.berlin.de/sen/uvk/verkehr/verkehrsplanung/radverkehr/weitere-radinfrastruktur/zaehlstellen-und-fahrradbarometer/

Senate Department for Urban Development, Building and Housing. (n.d.). Lebensweltlich orientierte Räume (LOR) Berlin [Living Environment Oriented Rooms (LOR) in Berlin] Shapefiles. , *Accessed: 10.02.2023*. Retrieved from https://www.berlin.de/sen/sbw/stadtdaten/stadtwissen/sozialraumorientierte-planungsgrundlagen/lebensweltlich-orientierte-raeume/

Strauss, J., & Miranda-Moreno, L. F. (2013, August). Spatial modeling of bicycle activity at signalized intersections. *Journal of Transport and Land Use*, *6*(2), 47–58. Retrieved 2023-08-30, from https://jtlu.org/index.php/jtlu/article/view/296 doi: 10.5198/jtlu.v6i2.296

Strava Metro. (n.d.). *Strava Metro - Berlin Data.* Retrieved 2022-10-05, from https://metro.strava.com/

Woodcock, J., Edwards, P., Tonne, C., Armstrong, B. G., Ashiru, O., Banister, D., ... Roberts, I. (2009). Public health benefits of strategies to reduce greenhouse-gas emissions: urban land transport. *The Lancet*, *374*(9705), 1930–1943. Retrieved 2022-03-21, from https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(09)61714-1/fulltext doi: https://doi.org/10.1016/S0140-6736(09)61714-1

Yi, Z., Liu, X. C., Markovic, N., & Phillips, J. (2021). Inferencing hourly traffic volume using data-driven machine learning and graph theory. *Computers, Environment and Urban Systems*, *85*, 101548. Retrieved 2022-03-01, from https://www.sciencedirect.com/science/article/pii/S0198971520302817 doi: https://doi.org/10.1016/j.compenvurbsys.2020.101548

Zahedian, S., Sekuła, P., Nohekhan, A., & 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. Retrieved from https://journals.sagepub.com/doi/full/10.1177/0361198120910737?casa_token=GWfT3xRIY3oAAAAA%3ATNCNkfF9kWCQKVrH2k43s9j9IBAQc2IoupPSleggGx5EO2jsOug6_HEJP7ir2khjiwMKoVT2ppZd (Publisher: SAGE Publications Sage CA: Los Angeles, CA) doi: https://doi.org/10.1177/0361198120910737