PREDICTING CYCLING TRAFFIC IN CITIES: IS BIKE-SHARING DATA REPRESENTATIVE OF THE CYCLING VOLUME?

Silke K. Kaiser
Data Science Lab
Hertie School
s.kaiser@phd.hertie-school.de

Nadja Klein
Research Center Trustworthy Data Science and Security
UA Ruhr,
Department of Statistics
Technische Universität Dortmund
nadja.klein@statistik.tu-dortmund.de

Lynn H. Kaack
Data Science Lab
Hertie School
kaack@hertie-school.org

ABSTRACT

A higher share of cycling in cities can lead to a reduction in greenhouse gas emissions, a decrease in noise pollution, and personal health benefits. Data-driven approaches to planning new infrastructure to promote cycling are rare, mainly because data on cycling volume are only available selectively. By leveraging new and more granular data sources, we predict bicycle count measurements in Berlin, using data from free-floating bike-sharing systems in addition to weather, vacation, infrastructure, and socioeconomic indicators. To reach a high prediction accuracy given the diverse data, we make use of machine learning techniques. Our goal is to ultimately predict traffic volume on all streets beyond those with counters and to understand the variance in feature importance across time and space. Results indicate that bike-sharing data are valuable to improve the predictive performance, especially in cases with high outliers, and help generalize the models to new locations.

1 INTRODUCTION

Promoting bicycle mobility in cities has several benefits: Cycling contributes to individual (Oja et al., 2011) and public health (Woodcock et al., 2009). However, more importantly, already a shift of 5% of vehicle kilometers from cars to bicycles would reduce transport-related greenhouse gas emissions by 0.4% (Lindsay et al., 2011). More detailed information on current bicycle use is needed for promoting cycling through infrastructure changes, such as new bike lanes or adjusted traffic light phases, in an evidence-based, locally and timely targeted manner (Heesch & Langdon, 2016), especially given scarce resources, or when assessing accident cycling safety (Strauss et al., 2015). Given the spatiotemporal complexity of cycling data, machine learning (ML) algorithms have created many new research opportunities in this context (Klemmer et al., 2018; Rolnick et al., 2023; Xie et al., 2020).

Long-term stationary bicycle counters provide data on sub-hourly, hourly, daily, or annual traffic volumes. As such counters are costly, they are only installed in limited numbers across a road network (Romanillos et al., 2016). While also often complemented with short-term (manual) counts (Ryus et al., 2014), the availability of bicycle traffic counts remains temporally and spatially limited. Accordingly, understanding how the observed counter measurements can be predicted to interpolate traffic volume beyond those eventually is key. In the literature, counter-measurements are so far analyzed through two main approaches. Firstly, studies detect unreliable data and interpolation missing entries (Beitel et al., 2018). Secondly, research focusing on the contributing factors behind

the counter measurements, including simultaneous or lagged weather conditions and time factors, such as the day of the week and the season, (Miranda-Moreno & Nosal, 2011), public and school holidays Holmgren et al. (2017), and bike-sharing and app-based data in predicting counter measurements Miah et al. (2022). However, it has yet to be explored how the various features used for the prediction of the counters, but especially the number of bike-sharing rides, vary in their feature importance across time and between counters. As Yi et al. (2021) argues, this knowledge is of importance for traffic planners when they want to predict traffic volumes under financial constraints. It allows prioritization of which data should be acquired, given their cost. If a higher prediction accuracy is required at specific times or places, this also gives an indication of which further indicators should be obtained for the given time or location.

Therefore, we first implement and benchmark various ML algorithms for predicting the hourly counts of stationary counters within a city, using inputs that have been shown to be pertinent in previous studies. The models are evaluated using RMSE. Secondly perform further experiments to explore how the feature importance for various inputs varies over time and space. Based on these findings, it shall thirdly be evaluated whether bike-sharing data would permit interpolating the predictions to generate city-wide estimates of cycling volume on the hourly and street level.

2 Data and Methods

2.1 Data

We use the city of Berlin as a case study. Bicycle counter data is available for 19 long and 12 short-term counter locations at the hourly level (SenUMVK). The dockless bike sharing data comprises nine months from 2019 for the providers Nextbike and Call-a-bike (CityLab Berlin). Additionally, we scrap the equivalent data for seven months in 2022 from Nextbike (Nextbike, 2020). The data contains the departure and arrival points and times of bike trips. We interpolate the trajectories using the bicycle routing algorithm from OpenStreetMaps (OpenStreetMap contributors, 2017). From the data, we exclude trips shorter than 100m and longer than 50km as well as trips shorter than 90 seconds and longer than 10 hours. In addition to cycling data, we use the hour, the weekday, and the month as features, data on public and school holidays (SenBJF), weather indicators (precipitation, sun, snow, temperature, wind speed, humidity, pressure at both the hourly and daily level) (meteostat), infrastructure information indicating the maximum speed, the type of bike lanes, the number of industry/shops/education within a 1km radius to the counter, and the distance to the city center (OpenStreetMap contributors, 2017), as well as socioeconomic indicators, such as the population density, average age and distribution of gender within the surrounding planning area. These planning areas are defined and used by the city for urban planning and are, on average, around 2 km^2 in size (Amt für Statistik Berlin-Brandenburg, 2020).

2.2 Methods

We predict the hourly long-term counter measurements using the above-stated data, treating each hourly observation individually. We use feature engineering for the bike-sharing data to create six features: The total count of bikes each taken, returned, or rented during a given hour in the city. Following Miah et al. (2022), we also create features of the same counts but within a 1km radius around the counter. Several algorithms are employed to predict the counter measurements, including multi-linear regression, regression tree, random forest, gradient boosting, XGBoost, support vector regression, and a shallow neural network. The performance is evaluated with the root mean squared error (RMSE). We validate using 10-fold cross-validation (leaving out a random set of hours) and with leave-one-group-out cross-validation using only the permanent counters (leaving each counter out once). The short-term counter serves as out-of-sample testing data.

In order to evaluate how the importance of bike-sharing features, and others, vary over time and space, we conduct several experiments. Firstly, we want to control how features vary in importance over time. We hypothesize, that bike-sharing might follow different usage patterns than other bike traffic; for example, because bike-sharing users differ from average cyclists, such as that they do not ride for leisure (Buck et al., 2013). Therefore, we train 24 separate models, one for each hour of the day. We repeat the same experiment training seven models, one for each weekday. Secondly, we want to control how feature importance varies over space. We hypothesize that bike-sharing is

used predominantly in the city center (Fishman et al., 2013). Therefore we train models separately, grouping counters by the distance to the city center.

3 RESULTS

	On Train		On Short-Term	
	CV	LOGO	CV	LOGO
Lin. Reg.	221.25	354.27	635.19	635.19
Dec. T.	144.10	163.15	255.89	224.28
RF	118.37	142.70	224.65	218.43
SVM	143.53	154.41	173.42	176.80
Grad. Boo.	51.50	142.07	253.42	223.50
XGBoost	62.16	147.07	213.36	257.30
NN	106.46	184.85	192.50	198.15

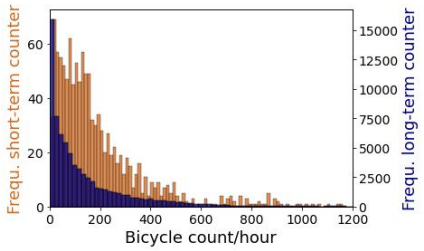

Figure 1: Predictive RMSE for various models, trained via 10-fold cross-validation and leave-one-group-out cross-validation. Both models are also evaluated via RMSE on the short-term counters as testing data. The histogram depicts the hourly long-term and short-term counter measurements.

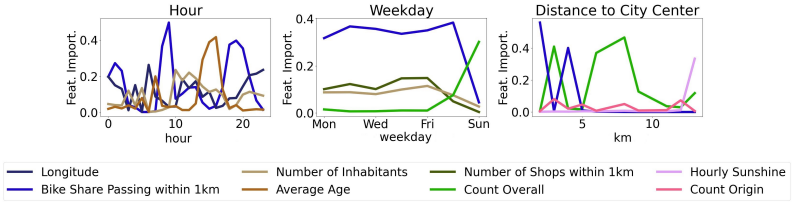

Figure 2: Feature importance across the hours, weekdays, and distance to the city center based on the random forest models which were hypertuned with LOGO cross-validation.

Figure 1 shows the predictive RMSE of models hypertuned and evaluated with a 10-fold cross-validation (CV) and a leave-one-group-out (LOGO) cross-validation, with the latter splitting the data such that each training set comprises all data except those belonging to one held-out counter of the 19 permanent counting stations. The table also reports RMSEs of the model tested on the short-term counters. The models perform reasonably well in the regular CV, but the errors remain high in the LOGO evaluation, which requires generalization to new locations. Further analysis shows that this is due to the models not capturing strong positive outliers well at specific counters. Figure 2 depicts each of the four most important features and their importance score of the random forest models trained separately for every hour, weekday, and distance to the city center. It becomes apparent that the count of passing bike-sharing riders within a 1km radius is the most important feature across days and during the morning commute. On Sundays and around 6 to 9 km from the city center, the city's overall count of bike-sharing users is the most important feature.

4 CONCLUSION AND FUTURE WORK

By providing data-driven insights into urban bike traffic volumes, we aim to aid infrastructure planning to promote sustainable urban mobility. Here, we focused on the importance of various features, in particular, on novel bike-sharing data, for the prediction of bicycle counts in order to enable a city-wide time-varying interpolation of bicycle volume at the street level. We propose to use our case study to evaluate how valuable different data sources are for this aim, in particular high-resolution bike-sharing data. Our preliminary results indicate that bike-sharing usage is an important feature for improving the prediction during commuting hours, which is also when the models tend to perform the worst due to high outliers.

In our future work, we plan to better understand and predict bicycle counts by evaluating the models at the daily level (Miah et al., 2022) and further analysis of spatial feature importance variation. Also,

further features indicating stationary similarities among the counters shall be included to possibly improve cross-location prediction. Additionally, we will conduct explainable time-series modeling and resulting feature importance analysis to control for temporal local dependencies (Afrin & Yodo, 2022). For example, this could give hints if short-term counts could be used in a temporally and spatially targeted way to improve prediction models.

We expect that the results of such analysis can be used to create temporally and spatially resolved maps of bike traffic counts. Moreover, they may provide insights into which data sources are most valuable for such predictions and allow cities to prioritize data collection. We aim for the results to inform urban infrastructure planning for net zero mobility in Berlin, and the approach and insights generalize beyond the case study.

Acknowledgments and Disclosure of Funding

This work has received funding from the European Union's Horizon Europe research and innovation programme 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). We also want to thank CityLab Berlin for providing us with the bike-sharing data for 2019.

References

Tanzina Afrin and Nita Yodo. A Long Short-Term Memory-based correlated traffic data prediction framework. *Knowledge-Based Systems*, 237:107755, 2022.

Amt für Statistik Berlin-Brandenburg. *Kommunalatlas Berlin*. Steinstraße 104–106, 14480 Potsdam, 2020. URL https://instantatlas.statistik-berlin-brandenburg.de/instantatlas/interaktivekarten/kommunalatlas/atlas.html.

David Beitel, Spencer McNee, Fraser McLaughlin, and Luis F. Miranda-Moreno. Automated Validation and Interpolation of Long-Duration Bicycle Counting Data. *Transportation Research Record*, 2672(43):75–86, 2018.

Darren Buck, Ralph Buehler, Patricia Happ, Bradley Rawls, Payton Chung, and Natalie Borecki. Are bikeshare users different from regular cyclists? A first look at short-term users, annual members, and area cyclists in the Washington, DC, region. *Transportation research record*, 2387(1):112–119, 2013.

CityLab Berlin. Shared Mobility Flows. Retrieved: 01.03.2022. URL https://bikesharing.citylab-berlin.org/.

Elliot Fishman, Simon Washington, and Narelle Haworth. Bike Share: A Synthesis of the Literature. *Transport Reviews*, 33(2):148–165, 2013.

Kristiann C. Heesch and Michael Langdon. The usefulness of GPS bicycle tracking data for evaluating the impact of infrastructure change on cycling behaviour. *Health Promotion Journal of Australia*, 27(3):222–229, 2016.

Johan Holmgren, Sebastian Aspegren, and Jonas Dahlströma. Prediction of bicycle counter data using regression. *Procedia Computer Science*, 113:502–507, 2017.

Konstantin Klemmer, Tobias Brandt, and Stephen Jarvis. Isolating the effect of cycling on local business environments in London. *PLOS ONE*, 13(12):e0209090, December 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0209090. URL https://dx.plos.org/10.1371/journal.pone.0209090.

Graeme Lindsay, Alexandra Macmillan, and Alistair Woodward. Moving urban trips from cars to bicycles: impact on health and emissions. *Australian and New Zealand Journal of Public Health*, 35(1):54–60, February 2011.

meteostat. The Weather's Record Keeper. *Retrieved: 01.08.2022*. URL https://meteostat.net/en/.

Md Mintu Miah, Kate Kyung Hyun, Stephen P. Mattingly, and Hannan Khan. Estimation of daily bicycle traffic using machine and deep learning techniques. *Transportation*, April 2022.

Luis F. Miranda-Moreno and Thomas Nosal. Weather or Not to Cycle: Temporal Trends and Impact of Weather on Cycling in an Urban Environment. *Transportation Research Record*, 2247(1): 42–52, January 2011.

Nextbike. Official Nextbike API Documentation, 2020. URL `https://github.com/nextbike/api-doc`.

P. Oja, S. Titze, A. Bauman, B. de Geus, P. Krenn, B. Reger-Nash, and T. Kohlberger. Health benefits of cycling: a systematic review. *Scandinavian Journal of Medicine & Science in Sports*, 21(4): 496–509, 2011.

OpenStreetMap contributors. *Planet dump retrieved from https://planet.osm.org*. 2017. URL `https://www.openstreetmap.org`.

David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. Tackling Climate Change with Machine Learning. *ACM Computing Surveys*, 55 (2):1–96, March 2023.

Gustavo Romanillos, Martin Zaltz Austwick, Dick Ettema, and Joost De Kruijf. Big Data and Cycling. *Transport Reviews*, 36(1):114–133, 2016.

Paul Ryus, Erin Ferguson, Kelly M. Laustsen, Robert J. Schneider, Frank R. Proulx, Tony Hull, Luis Miranda-Moreno, National Cooperative Highway Research Program, Transportation Research Board, and National Academies of Sciences, Engineering, and Medicine. *Guidebook on Pedestrian and Bicycle Volume Data Collection*. Transportation Research Board, Washington, D.C., 2014.

SenBJF. Ferientermine [Vacation Dates]. Retrieved: 02.08.2022. URL `https://www.berlin.de/sen/bjf/service/kalender/ferien/artikel.420979.php`.

SenUMVK. Zählstellen und Fahrradbarometer: Fahrradverkehr in Zahlen [Counting stations and bicycle barometer: Bicycle traffic in figures]. Retrieved: 18.02.2022. URL `https://www.berlin.de/sen/uvk/verkehr/verkehrsplanung/radverkehr/weitere-radinfrastruktur/zaehlstellen-und-fahrradbarometer/`.

Jillian Strauss, Luis F. Miranda-Moreno, and Patrick Morency. Mapping cyclist activity and injury risk in a network combining smartphone GPS data and bicycle counts. *Accident Analysis & Prevention*, 83:132–142, 2015.

James Woodcock, Phil Edwards, Cathryn Tonne, Ben G Armstrong, Olu Ashiru, David Banister, Sean Beevers, Zaid Chalabi, Zohir Chowdhury, Aaron Cohen, Oscar H Franco, Andy Haines, Robin Hickman, Graeme Lindsay, Ishaan Mittal, Dinesh Mohan, Geetam Tiwari, Alistair Woodward, and Ian Roberts. Public health benefits of strategies to reduce greenhouse-gas emissions: urban land transport. *The Lancet*, 374(9705):1930–1943, 2009.

Peng Xie, Tianrui Li, Jia Liu, Shengdong Du, Xin Yang, and Junbo Zhang. Urban flows prediction from spatial-temporal data using machine learning: A survey. *Information Fusion*, 59:1–12, 2020.

Zhiyan Yi, Xiaoyue Cathy Liu, Nikola Markovic, and Jeff Phillips. Inferencing hourly traffic volume using data-driven machine learning and graph theory. *Computers, Environment and Urban Systems*, 85:101548, 2021.