**BRUBIKE: A Dataset of Bicycle Traffic and Weather Conditions for Predicting Cycling Flow**

Steven Vanden Broucke†, Luis Manuel Vidal Piña‡, Tien Huu Do†, Nikos Deligiannis†

†Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium
‡Universidad de las Ciencias Informáticas, Havana, Cuba
Email: {svandenb, lvidalpi, thdo, ndeligia}@ietovub.be

**Abstract**—Based on historical bike counting information, geographical and temporal patterns in human mobility can be detected. Predicting bicycle traffic and traveler flows enable the identification and prevention of potential bottlenecks in a city’s cycling network and creates new opportunities for mobility solutions. Since the introduction of the first bicycle counting station in Brussels in 2017, the city has expanded its counting network to twelve stations and is aiming to reach fifteen stations by the end of 2019. Real-time and historical bike counting data concerning these stations is made available to the public through web endpoints. In this study, we introduce BRUBIKE, a novel aggregated dataset of bicycle and meteorological information concerning the city of Brussels. We aim to lower the boundary of accessing Brussels’ cycling information and to stimulate the creation and evaluation of novel traffic flow models on Brussels’ data. A subset of existing machine learning models is evaluated on the proposed dataset with the task of predicting bicycle traffic for a yet unseen period, once with weather parameters, and once without weather parameters. Results indicate significantly better prediction performance when weather parameters are included due to the existing correlation of weather and bike traffic. Finally, we propose an open source application to make historical bike traffic and predictions more accessible towards Brussels’ citizens.

**Index Terms**—Data visualization, Big Data applications, Machine Learning, Internet of Things

**I. INTRODUCTION**

In recent years, increasing concerns about global warming, environmental issues, gasoline prices, the urban noise problem and air pollution have stimulated the modern society to shift towards a more sustainable transport model [1], [2]. Cycling mobility is one of the transportation modalities that is being actively promoted in many cities around the world [3]. In Belgium, bicycle use has steadily increased and became one of the most important means of transportation in 2015 for approximately one third of Belgian citizens [4].

The pedal electric cycle, formally known as pedelec, enables commuters to travel to an assisted speed up to 25km/h. In 2015, pedelecs or e-bikes accounted for 31% of all bicycle sales [5] in Belgium, this number further grew to 39% [6] in 2016, 45% in 2017 [7], and nearing 50% in 2018 [8]. However, due to the limited speed and range of these pedelecs, commuters that have to travel longer distances were not yet convinced to switch from their traditional form of transport. The introduction of 45km/h capable speed pedelecs, formally known as speedelecs, gave an answer to some of the existing drawbacks of the traditional bicycles and pedelecs. Due to higher speeds, longer ranges, and comparable travel times to public transport or car, more commuters felt speedelecs to be a viable alternative transportation mode [6]. Other contributors to the growth of bicycle traffic in Brussels are bike sharing programs and food deliveries by bike. Since the introduction of the first docked bike sharing system by the Villo! bicycle rental program in 2009, the number of bicycle docking stations in Brussels has grown to 352. Next to docked rental programs, free-floating bike rental programs are being introduced and regulated. However, in order to sustain this growing cycling modality in Brussels, understanding and predicting geographical and temporal patterns in bike traffic is a key objective [9].

In order to capture the growth in bicycle traffic, Brussels has installed twelve counting poles strategically positioned around the city. The poles display the number of cyclists that passed the pole during the current day and whole year. In this way the city tries to sensitize and motivate bicycle usage towards its citizens. With a similar philosophy, the city publishes real-time and historical bicycle information concerning their poles as open data¹. This published information does not only facilitate the process of urban planning and road engineering in the city, but could also potentially stimulate the creation of new businesses that focus on cycling mobility. Furthermore, data concerning long-term increases in bicycle traffic can help validate the city’s effort to promote and stimulate the shift towards cycling as a means of transportation.

Next to the geographical and temporal importance in traffic flow, there also exists a relation between some meteorological properties such as humidity and temperature. It has been proven that in some scenarios, considering weather information parameters like temperature, humidity and precipitation for the prediction of vehicle traffic flow, can achieve up to 25% improved error rates [10]. However, this applies to vehicles and not bicycles. Related research to bicycle traffic in Brussels covers assessing the risk to air pollution [11], the risk of accidents [12] and psychosocial and environmental correlates of cycling [13]. Forecasting problems in related work are typically applied to demand forecasting. In the work of [2], the authors utilise classified

¹www.villo.be
²www.data-mobility.brussels/bike/api/counts/
support vector regression with particle swarm optimization to forecast bike demand and parking usage in Taipei. The authors of [14] propose a hierarchical prediction model to predict the number of bikes that will be rent or returned from a rental station, thus, reducing unbalanced stations. The published dataset by UCI [15], consists of bike rental counts combined with weather data which can be utilised to develop new forecasting models for rental demands, however, this does not directly apply to traffic flow forecasting. Gallop et al. [16] makes use of a seasonal autoregressive model of Vancouver bicycle traffic using weather variables related to forecast bike counts, which directly relates to the presented data in this study. Since the Brussels’ bike count history was only available starting from December 2018, no existing studies were found to exploit this recent data set. Furthermore, the majority of existing works handle rental demands and not traffic flow counts and cannot be directly applied to the presented problem.

In this research, three contributions are delivered:

- **BRUBIKE**, an aggregated dataset of bike traffic and weather information concerning Brussels’ is made available to the public
- A subset of existing machine-learning based models is evaluated on the task of predicting bike traffic on top of the BRUBIKE dataset.
- An open source platform for the exploration of Brussels’ cycling data is introduced, allowing citizens to explore historical, real-time and predictive bike data.

The remainder of this paper is structured as follows. We first describe the acquired data sources, how they are aggregated and how they are made available to the public. Section III details a subset of existing models and compares their performances on the newly acquired data set, with and without weather information. Finally, we conclude the paper with the description of an open-source platform that enables the analysis of Brussels bike traffic flow in Section IV.

### II. DESCRIPTION OF BRUBIKE

Concepts such as Smart Cities have evolved, where data generated by Internet of Things sensors are used for the analysis and prediction of air quality indices, traffic flow, noise and many others [17]. In order to create and evaluate traffic flow models, data from all cities around the world is required. BRUBIKE aims to make this data concerning the city of Brussels more accessible to other researchers.

#### A. The BRUBIKE dataset

BRUBIKE consists out of the historical bike counting data from Bruxelles Mobilité³ combined with web-scraped historical weather information⁴ of Uccle’s weather station in Belgium. The initial version of the BRUBIKE dataset is identified as BRUBIKE_V1_1819 and is utilised throughout this study. Since the collection of historical bike counting data from nine stations in the city, 3 new poles were introduced. However, due to a lack in historical data on these poles, these poles were not yet included into the dataset. The BRUBIKE dataset is made publicly available as a tab-separated values (TSV) file on Github⁵. In Figure 1, the BRUBIKE record captures Brussels’ historical bike counting and weather information in one format, which reduces complexity and promotes the evaluation of novel travel prediction models. The collected data spans over a period of 193 days from 6 December 2018

---

³ [www.data-mobility.brussels](http://www.data-mobility.brussels)
⁴ [www.timeandate.com](http://www.timeandate.com)
⁵ [www.github.com/vandenbroucke/BRUBIKE](http://www.github.com/vandenbroucke/BRUBIKE)

---

TABLE I
BRUBIKE PARAMETER DESCRIPTION

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>device_name</td>
<td>unique identifier for a bike counting pole</td>
</tr>
<tr>
<td>latitude</td>
<td>angular distance north or south of the earth’s equator</td>
</tr>
<tr>
<td>longitude</td>
<td>angular distance east or west of the Greenwich meridian</td>
</tr>
<tr>
<td>timestamp_from</td>
<td>start of counting interval in Unix epoch seconds</td>
</tr>
<tr>
<td>timestamp_until</td>
<td>end of counting interval in Unix epoch seconds</td>
</tr>
<tr>
<td>bike_count</td>
<td>number of bicycles counting during the counting interval</td>
</tr>
<tr>
<td>bike_avg_speed</td>
<td>average speed of all counted bicycles during the counting interval</td>
</tr>
<tr>
<td>weather_timestamp</td>
<td>time of weather measurement in Unix epoch seconds</td>
</tr>
<tr>
<td>temperature</td>
<td>temperature in degrees Celsius</td>
</tr>
<tr>
<td>weather_condition</td>
<td>type of weather (Sunny, Fog, Partly cloudy,...)</td>
</tr>
<tr>
<td>wind_speed</td>
<td>speed of wind in km/h</td>
</tr>
<tr>
<td>wind_direction</td>
<td>direction of the wind</td>
</tr>
<tr>
<td>humidity</td>
<td>percentage of water vapour present in the air</td>
</tr>
<tr>
<td>barometer</td>
<td>atmospheric pressure indicated in mbar</td>
</tr>
<tr>
<td>visibility</td>
<td>distance of visibility indicated in km</td>
</tr>
</tbody>
</table>
until 16 June 2019 and consists out of 102,310 records. In Table I the available parameters of each BRUBIKE record are described. Due to disruptions in service availability and works on infrastructure, some counting stations will not have taken measurements during these disruptions. Similarly, for the web scraped weather data, some parameters might be empty during specific periods.

Typically, the bike counting interval is fifteen minutes. However, as this parameter is prone to change, timestamps for the start and end are included in the BRUBIKE records. Due to different asynchronous measurement intervals for the bike counting and weather measurements, the closest weather_timestamp is matched to a record’s bike counting interval (timestamp_from to timestamp_until).

**B. Data Correlation**

In Figure 2, the Pearson correlation coefficients between a subset of weather parameters and each bike counting station’s bike_count is given. In order to validate correlations between the textual weather types and the bike count, all unique types of weather conditions are first extracted and added as a binary feature vector. Each BRUBIKE record will then contain an additional set of 33 binary features for each possible weather type. If the weather was sunny for that record, a value of 1 will be observed in the feature ‘sunny’. For sake of clarity, features with a close to zero correlation coefficient were ignored in the figure. For a selected station, the Pearson correlation coefficient [18] of $X$, all bike_count values for that station and $Y$, the selected weather parameter values, is: 

$$
\rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}
$$

where $\text{cov}$ is the covariance, $\sigma_x$ is the standard deviation of $X$, and $\sigma_y$ is the standard deviation of $Y$. Based upon the presented correlation coefficients in Figure 2, we can see that, for the majority of counting stations, there exists a positive correlation for temperature and bikes counted, and a negative correlation for humidity and bikes counted. This naturally indicates that there are more cyclists when it is warm and less cyclists when it is wet. However, for some areas there exists close to no correlation between the weather parameters and bike counts. Similarly, from the presented weather types in the weather condition parameter, we find small correlations coefficients with the bike count. Weather conditions correlations are not as significant as temperature and humidity; however, they can still be exploited to improve traffic predictions. It is important to note that for some bike counting stations the weather features might not improve a model’s performance due to close to zero correlation coefficients. This indicates that for some locations, the amount of cyclists is almost not affected by the weather. Due to the relatively young age of the dataset, it does not span over a full year. A future extended version of the dataset could enable us to better capture seasonal patterns and correlations with the measured bike counts, and improve bike traffic flow predictions.

**III. BRUBIKE, BENCHMARK EVALUATION**

In this section, we discuss how the BRUBIKE parameters are formatted in order to be consumed by a variety of machine-learning models. An existing subset of machine-learning models is listed and validated with the task of predicting bike traffic for an unseen period in time, once with and once without weather parameters for that period.

**A. Experimental Settings**

BRUBIKE records are preprocessed before being fed to a set of machine-learning models. As the weather condition parameter can be considered as a categorical variables, and machine learning algorithms require numerical input, the parameter is converted using one-hot encoding to a vector of 33 binary features. On top of these new features, the timestamp_from is converted to two additional features: hour_of_day_slice (0 to 23) and day_of_week_slice (0 to 6). After this conversion, features with a close to zero correlation coefficient to the bike count are discarded. In Table II, the selected features with a correlation coefficient of $|\rho| \geq 0.1$ are depicted. Finally, outliers concerning bike counts are removed using the Interquantile Range Rule (IQR) with $q_1 = 0.25$ and $q_3 = 0.75$. This results in a removal of 719 records, which is 7.016 % of the BRUBIKE dataset. In Figure 3, a box plot containing the quantile distributions of the bike_count parameter is depicted. The remaining dataset contains approximately 94K records. The dataset is split into a training set, containing 80% of the BRUBIKE dataset, which is the first 144 days. The left over 20% consists out of the last 36 days and is used for testing.

We employed four machine learning models for predicting the number of bikes: a support vector machine regression (SVR) [19], a random forest regression model (RFR) [20],
an extreme gradient boosting model (XGBoost) [21] and a fully-connected neural network model (FNN) [22]. We fine-tuned the parameters of the considered models using grid search. We employed scikit-learn\(^6\) to implement the SVR and RFR models. The gradient boosting model is implemented using XGBoost and we use Tensorflow\(^7\) for the FNN model.

For the SVR model, we employed radial basis function (rbf) as the kernel. In addition, the degree of the polynomial kernel function and the epsilon value were set to 3 and 0.1, respectively. For RFR, the selected value for the maximum depth of the tree (max_depth) was 7 and the selected number of trees in the forest (n_estimators) was 8. For XGBoost the selected value for learning rate was 0.1, selected value for the maximum depth (max_depth) was 5, the alpha value for L1 regularization was 10 and the number of trees in the XGBoost model (represented by the n_estimators parameter) was 300. For the FNN model, we used two hidden layers; the first hidden layer has 28 nodes and the second layer contains 14 nodes. The ReLU function was used as the activation function. To train the FNN, we selected the mean square error (mse) as the objective function and optimized it using the Adam algorithm [23] with a learning rate of 0.001. To handle overfitting, we used classical weight decay regularization with a regularization rate of 0.0001.

In order to measure the performance of the considered models, we employed root mean square error (RMSE) and mean absolute error (MAE) as the evaluation metrics. To achieve robust results, we ran each considered model 5 times and calculated the average of the output scores. Additionally, in order to see the impact of the weather conditions on the amount of bikes, we inspect the performance of the considered models using all the features, and compare the performance without the weather features.

\(^6\)https://scikit-learn.org/stable/

\(^7\)www.tensorflow.org

B. Experimental results

We outline our results in Table III. First, we compare various models using all included features (locations, hour of the day, day of the week, and weather conditions as humidity, temperature, barometer, visibility,..). It is clear that the RMSE is between 4 and 6.5, and the MAE is between 3.4 and 4.5. For both performance metrics, the FNN delivers the best results.

For the FNN model, we observe a 12% reduction in RMSE, thus, indicating a clear improvement for predicting bike counts with weather information. When weather features were excluded, the performances of all the considered models were worse. Specifically, the RMSE score is between 5.4 and 6.4 and the MAE score is between 3.7 and 4.5. Furthermore, the RFR model performs the best among the others. The reason is that the FNN, SVR and XGBoost models tend to overfit more when the number of considered features is limited. A common weakness of RFR models is that they don’t tend to fit well for increasing and decreasing trends. Thus, RFR might perform worse for future larger dataset that spans over all weather seasons, or multiple years.

On Figure 4, we show the Ground truth (green) trend for one week on bike station CJM90 (located in the crossroad between Quai des Charbonagges street and Rue Vandermaelen), and the predicted values for the two best models on the benchmark evaluation: FNN (blue) and XGBoost (pink). As can be seen, the predicted values follow the Ground truth across the time

\[^{6}\]https://scikit-learn.org/stable/

\[^{7}\]www.tensorflow.org
axis. Specifically, the trained models predict more bikes on weekdays compared to weekends. Furthermore, the predicted amount of bikes during day-time is much higher than during night-time. In general, the trained models are able to capture the seasonal pattern of the BRUBIKE dataset.

IV. WEB APPLICATION

Existing open data platforms that publish Brussels’ bike traffic data do not effectively sensitize its users on the bicycle traffic situation in the city. Typically, historical and real-time bike counting data is made available in a non-visual raw format. Furthermore, station locations are indicated on a map using static markers without any options for user interaction. With the limited tools that are provided by these platforms, users are unable to quickly grasp the current and passed bicycle traffic situation in their city. In order to more effectively sensitize Brussels’ citizens on the bicycle traffic, we developed an open source web application which allows for its users to explore Brussels bike traffic situation in an interactive fashion. The application is made publicly available on Github.

In Figure 5, an initial version of the proposed web application is shown. The application consists out of a dynamic map (a) of Brussels, where bike stations are visualized using dynamic circle markers. Upon interaction with the provided timeline (b), these markers will change their content, color and size dynamically, depending on the bikes counted at that station on the selected point in time. Upon the selection of a counting station, one is able to explore historical bike counting information in a graph format (c).

V. CONCLUSION AND FUTURE WORK

In this study, historical bicycle traffic information and meteorological information concerning the city of Brussels was collected, processed and made available through the BRUBIKE dataset. The lack of an existing dataset concerning bike traffic and weather in Brussels motivated the authors to develop BRUBIKE. The proposed dataset offers a less complex data format compared to the readily available web endpoints and reduces the time required to validate novel traffic flow prediction models. As no other works exploited this new data, BRUBIKE is the first of its kind for Brussels. Furthermore, existing models developed for other cities can now be tested against unseen data from another geographic location. Correlations between the added weather parameters and bike counts were found and exploited by a set of machine-learning models. Results indicate an improved performance for three of four models for predicting bike traffic in an unseen period when weather parameters are included. Finally, an interactive web application is introduced to sensitize Brussels’ citizens on the historical bicycle traffic in their city.

As Brussels’ bike counting network is expanding, we look to append the historical information of these newly added counting stations in a future version of the BRUBIKE dataset. Additionally, we aim to publish a new version of the BRUBIKE dataset once it spans over a full period of one year, and is able to capture all seasonal patterns. We acknowledge the existence of autoregressive models such as the work by Gallop et al. [16] for the prediction of traffic flow and plan to experiment with these models on the BRUBIKE dataset as future work. Once a well performing model is selected, we plan to integrate it into the proposed web application. In this way, users of the web application will be able to see predicted bike counts for coming periods and their relation to the weather forecast.

---

Fig. 5. The BRUBIKE interactive web application for the exploration of bike traffic in Brussels.
ACKNOWLEDGEMENT

This work is financially supported by Innoviris\textsuperscript{10} (Brussels Institute for Research and Innovation, MobiWave project). L. M. V. Piña acknowledges the financial support of VUB-Cuba PhD scholarship.

REFERENCES


\textsuperscript{10}http://www.innoviris.be/