Study of the correlation between the weather and the demand for bicycles in GIRA stations

Ana Santos (n.84364)[1]

[1] *Departamento de Física, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal*

The aim of this paper is to study the impact of weather on bicycle demand. To this end, data from the public bike sharing system in Lisbon (GIRA) and weather station were explored. A methodology was proposed to preparation the data and to calculate the correlation with the removal of intraday and interday seasonality in the demand for bicycles. It was found that the bicycle use in the area studied was mostly for commutes, and that the data as daily and weakly periodicity. It was also found, from Pearson correlations that the demand is positively (yet softly) correlated to the temperature and negatively (yet softly) correlated to the wind intensity. No delineated correlations were found for the precipitation and humidity. DCCA analysis confirms the present of non-linear and seasonal aspects associated with demand, and DCCA-l suggests that check-ins and wind intensity are correlated.

I. INTRODUCTION

The ILU project or Integrative Learning from Urban Data and Situational Context for City Mobility Optimization unites INESC-ID, Câmara Municipal de Lisboa and LNEC (Laboratório Nacional de Engenharia Civil), in a joint effort to address two major challenges: 1) the lack of an integrative analysis capable of combining different sources of urban data collected from city sensors and ticket validations in the diverse modalities of public transport, and 2) the absence of situational context in predictions and public transport planning.

This work is inserted in this project, focusing on the bicycle mobility in particular the public bike sharing system in Libon (GIRA). The main objectives of this work are: 1) the exploratory analysis of the GIRA's data with the aim of finding a well defined group of bicycle stations and time period to use in the rest of the work, 2) consolidate the date collected from the public bicycle stations with the data from weather stations and 3) study the correlation between the data sources in order to understand thee effect of the weather on the demand for bicycles.

II. BACKGROUND

A. Bike sharing systems

A bike sharing system is a service in which bicycles are made available for shared used by individuals for a short term basis. The bicycles are stored in docks in different stations. The bicycles can be removed from a dock (check-in) and then return to a different dock and station, as long as it belongs to the same system (check-out). The number of check-ins and check-outs give a measure of the demand for bicycles at a given time and location. The true demand can be masked when these stations are full or empty (people looking get and drop a bike respectively would not be able to do so).

B. Time-Series

A time series x_i is a series of data points ordered in time, $x = x_1, x_2, x_3, ...x_T$. Normally the data points are given for equally spaced points in time.

C. Auto-Correlation

The auto-correlation measures how a time series is related with its past. In other words, it measures the correlation between a time series $x = x_1, x_2, ...x_{T-k}$ and the same time series shifted by a time lag k, $x = x_k, x_{k+1}, ...x_T$. The expression for the auto-correlation is:

$$r_k = \frac{\Sigma_{i=1}^{T-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\Sigma_{i=1}^{T} (x_i - \bar{x})^2} \quad . \tag{1}$$

D. Pearson-Correlation (PCC)

The Pearson correlation coefficient (PCC) is a measure of the linear correlation between two variables time series, x and y, given by:

$$r_{x,y} = \frac{\Sigma_{i=1}^{T} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{T} (x_i - \bar{x})^2 \Sigma_{i=1}^{T} (y_i - \bar{y})^2}} \quad . \tag{2}$$

The Pearson coefficient is invariant under linear transformations of the two variables. PCC's of values -1 or 1 imply that a linear equation describes the relationship between x and y perfectly. 0 implies that there is no linear correlation.

E. Detrended Cross-Correlation Analysis (DCCA)

The previous statistics either assume time series stationarity (auto-correlation) or linearity on the relationship between time series (PCC). Boris Podobnik [1] presents an alternative statistic, DCCA to investigate power-law cross-correlations between time series in the presence of non-stationarity. In this method the time series x and x' of length T, are divided into T-n overlapping boxes, each containing n+1 values. Defining $R_k = \Sigma_{i=1}^{k} x_i$ where k=1,...N as the *integrated signal*, $\tilde{R}_{i,k}$, where $i \leq k \leq i+n$ as the local trend (given by the ordinate of the least square fit), and the *detrended walk* as the difference between the two, the covariance of the residuals in each box is calculated by:

$$f_{DCCA}^2(n,i) = \frac{1}{n-1} \Sigma_{k=i}^{i+n} (R_k - \tilde{R}_{i,k})(R'_k - \tilde{R}'_{i,k}) \quad . \tag{3}$$

And so the detrended covariance is given by:

$$F_{DCCA}^2(n) = \frac{1}{N-1} \Sigma_{i=1}^{T-n} f_{DCCA}^2(n,i) \quad . \tag{4}$$

If $(R_k = R'_k)$ the detrended covariance reduces to the detrended variance F_{DFA}^2 used in the detrended fluctuation analysis (DFA) method [2].

F. DCCA-l with periodic trends

The previous method does not work well with time series with periodic trends. Horvatic et al [3] proposed a DCCA variant with varying polynomial order l. In this method the detrended covariance is calculated by:

$$F^2_{DCCA}(n) = \frac{A_2B_2}{A_nB_n}\left(\Sigma_{i=1}^{N-n}\left|f^2_{DCCA}(n,i)\right|\right) \quad . \tag{5}$$

To calculate the local trend $\tilde{R}_{i,k}$ instead of using the ordinate of the least square fit, the considered value is the fit to a polynomial with order l, which increases with the size of the box. The term $\frac{A_2B_2}{A_nB_n}$ is added to normalize the covariance since l increases with n.

G. DCCA-l coefficient

The introduced methods do not properly quantify the level of cross-correlation. To solve this, Zebende [4] proposed a coefficient based on DFA and DCCA methods. This coefficient has values in [-1,1] and is defined as:

$$\sigma_{DCCa} = \frac{F^2_{DCCA}}{F_{DFA}\{y_i\}F_{DFA}\{y'_i\}} \quad . \tag{6}$$

III. RELATED WORK

Flyn, Dana, Sears and Aultman-Hall [5] used data from surveys about commuting in a northern U.S. state and a generalized linear model. They found that the likelihood of bicycle commuting increased in the absence of rain and with higher temperatures, and decreased with snow and wind. No effects were found for daylight hours.

Miranda-Moreno and Nosal [6] used data from automatic counters in utilitarian bike facilities in Montereal, Canada. Two models were developed: one for the absolute (log-linear model and count regression models) and other for the relative (linear model) ridership. They found that the wind speed has no impact on results,that humidity levels have a significant effect and that generally an increase in temperature increases the number of bike counts, but if the temperature is above 28°C and the relative humidity is above 60% it has the opposite effect. The occurrence of precipitation in the current hour, in the last 3 hours and in the morning reduces ridership.

Thomas, Jaarsma and Tutert [7] studied cycle flows from utilitarian and recreational paths in the Netherlands. A bi-level model for predicting the demand for cycling was used. They found that humidity and visibility have no impact, temperature contributes the most and has a positive impact, followed by the rain and that the wind contributes the least.

Gallop, Tse and Zhao [8] use data from automatic counters in the city of Vancouver. An autoregressive integrated moving average (ARIMA) was used to model and predict the bike . They found that the weather as a significant effect on bike traffic by that this effect is exaggerated on models that don't account for correlation patterns in the error terms. Rain both current and lagged up to 3 hours were accepted in the model, snow and fog were rejected. The increase of temperature has a significant positive effect, but no negative effect was found when the temperatures are high. The humidity and clearness are only marginally significant. They also made a survey, which concluded that cyclists base their decision to bike on current rather than forecasted weather.

El-Assi, Nahmoud and Habib [9], used data from a public bike share system in Toronto, Canada. They used multilevel/linear model and a distributed lag model. They adjusted the temperature for wind chill and humidity (perceived temperature). They found positive correlation between bike share activity and temperature increase, negative correlation with precipitation, snow and humidity.

Ashqar, Elhenawy and Rakha, [10], used data from a bike station, the number of bicycles in each station, in the San Francisco Bay Area Bike Share system and uses the random forest technique to rank the predictors that were used to develop a regression model using a guided forward step-wise regression approach. They found that time-of-day, temperature and humidity level are significant count predictors and that a given station in time and the time-of-day were the most significant variables in the estimation of bike counts. The precipitation was not a significant predictor.

IV. DATA PREPARATION AND ANALYSIS

The data collected by GIRA is a set of events, where each event corresponds to a bike park or removal on a given station along with the resulting available and occupied docks on that station. The GIRA stations chosen for the analysis were the stations with identifiers 406, 407, 408, 416 and 417, located in the Saldanha roundabouts and the area behind Instituto Superior Técnico, see figure 10 with the location of the stations. To understand the average number of check-ins and check-outs, time series were created from the available data considering 60 minutes of granularity. Two other time series for the check-outs and check-ins were a made using the sum of the variation of bikes in each time interval. The weather stations collect/capture data on temperature (°C), humidity (%), wind intensity (km/h) and accumulated precipitation (mm). The weather station chosen for the analysis was the one with identifier 579 located near the Lisbon airport due to to its proximity to selected stations and low rate of missing values. From the raw weather data, we produced weather time series with a granularity of 60 minutes.

The data was available for the months of January and February of 2019. To remove the effects on the bicycle demand due to the holidays the first week of January was removed.

In Figures 1, 2, 11, 12 and 3, the data from one of the studied weeks is presented. As can be seen from the figure 1 the number of bikes has daily periodicity, the number of docked bikes are at a minimum during the night, and it reaches the maximum in the middle of the day. It can also be seen that the number of bikes varies less in the weekends. As such the collected data also has weekly periodicity. This can be confirmed by observing figure 2: the peak for the check-ins are before the peak for the check-outs, and their absolute values are lower for the weekends. This implies that the majority of users use this station for the commutes to work (in this area). In Table IV the maximum, minimum, mean and standard deviation of the data from the target period (7/1/2019 to 28/2/2019) are presented.

The Pearson correlation of the check-out and check-in and the weather data was then calculated.

For the Pearson correlation analysis, the correlation due to the daily and weekly periodicity of the data was removed by using the average of the data for a period of two off-peak hours (from 11h to 13h and from 14h to 16h) for working days. Periods with missing data were not considered.

For the analysis of DCCA and DCCA-l statistics, since the boxes are overlapping and since these statics assume temporal

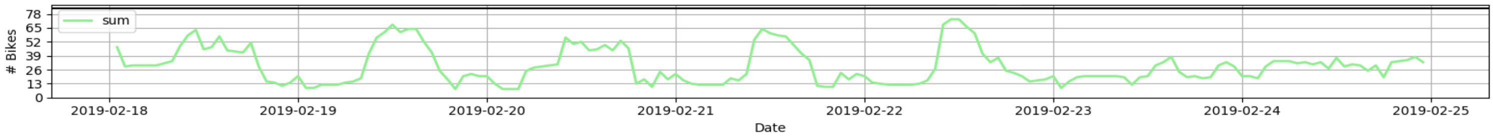

FIG. 1: Sum of docked bicycles in stations 406, 407, 408, 416, 417 from 18/2/2019 to 24/2/2019

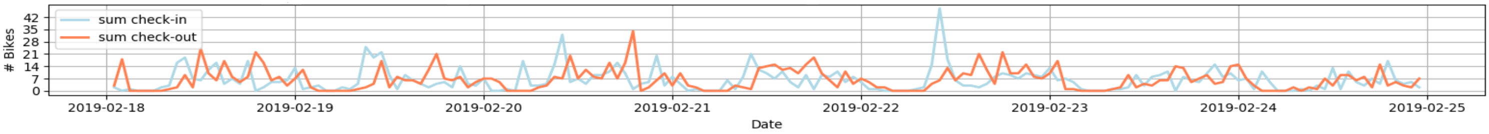

FIG. 2: Sum of check-ins and check-outs for stations 406, 407,408,416 and 417 from 18/2/2019 to 24/2/2019

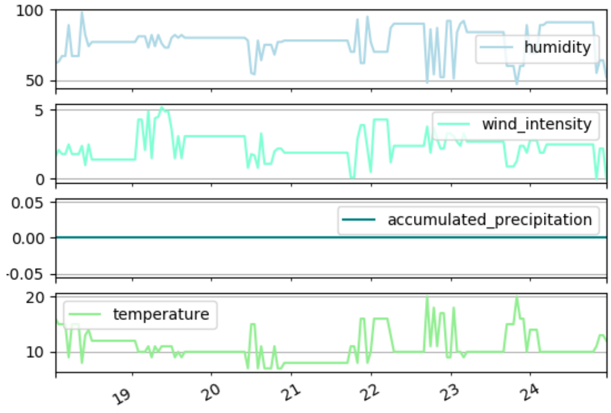

FIG. 3: Weather data from 18/2/2019 and 24/2/2019. Temperature is in °C, the accumulated precipitation is in mm, the wind intensity is in km/h and the humidity is in percentage

	station	max	min	mean	std
bikes	sum	79	1	28.82	14.97
check-in in 60 min	sum	87	0	5.74	8.59
check-out in 60 min	sum	87	0	5.75	8.62
temperature (°C)	579	20.0	4.0	10.56	3.33
acc. precipitation (mm)	579	3	0	0.02	0.19
wind intensity (km/h)	579	10.1	0	2.83	1.77
humidity (%)	579	98	33	71.73	13.19

TABLE I: Maximum, minimum, average and standard deviation for the sum of number of bikes, check-ins and check-outs in stations 406, 407,408,416 and 417 and for the parameters of weather station 579

continuity cutting out parts of the data would not work, so all daily periods were considered. In particular, we selected data pertaining to the month of February only since in January there was a period of 3 days with no data.

V. DISCUSSION AND RESULTS

A. Pearson correlation

The results are provided in Table V A an example of the scatter plot of the data (for the sum of check in and check out) is in figure 9. As can be seen for the target group of stations, the correlation between the sum of check-ins or check-outs and temperature is always positive and relatively small. The correlations with the precipitation do not agree on the signal, this might be due to the lack of representativity of variations on this variable since in the two monitored months, it only rained twice. The correlation for the wind is always negative and bigger than the others, and for the humidity does not agree on signals and is small. The scatter plots (Figure 9) suggest a lack of linear correlation, which agrees with the results.

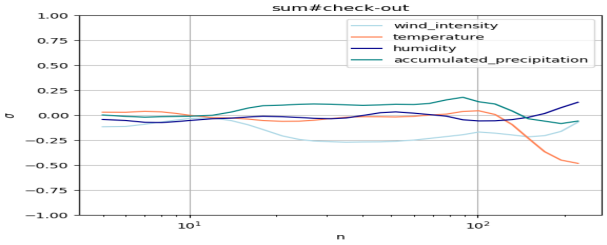

FIG. 4: DCCA coefficient for the check-outs.

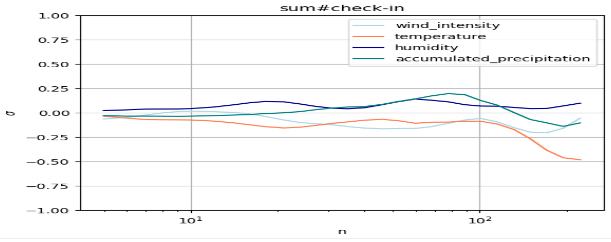

FIG. 5: DCCA coefficient for check-ins.

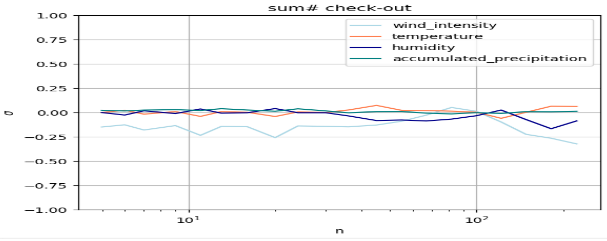

FIG. 6: DCCA-l coefficient for check-outs.

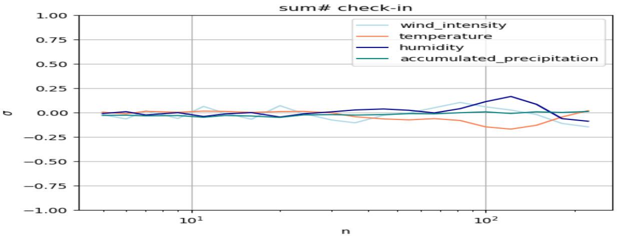

FIG. 7: DCCA-l coefficient for check-ins.

station			406	407	408	416	417	sum
temp	check-in	11-13	0.147	0.178	0.491	0.043	0.050	0.239
		14-16	0.127	0.255	0.050	0.050	0.088	0.138
	check-out	11-13	0.112	-0.171	0.273	0.190	-0.057	0.090
		14-16	0.303	0.082	-0.065	-0.065	0.115	0.167
prec	check-in	11-13	0.124	0.161	0.151	0.251	-0.070	0.161
		14-16	-0.204	0.017	0.005	-0.163	-0.011	-0.119
	check-out	11-13	-0.423	-0.146	-0.420	-0.124	-0.237	-0.414
		14-16	0.146	-0.344	-0.205	-0.267	0.287	-0.068
wind	check-in	11-13	-0.029	-0.044	-0.033	-0.248	-0.410	-0.288
		14-16	-0.122	-0.276	-0.116	-0.201	-0.251	-0.268
	check-out	11-13	-0.417	-0.412	-0.398	-0.147	-0.258	-0.501
		14-16	-0.140	-0.471	-0.404	-0.332	0.097	-0.337
hum	check-in	11-13	0.067	0.278	0.235	-0.112	-0.008	0.111
		14-16	0.080	0.111	0.027	0.058	0.081	0.100
	check-out	11-13	-0.107	0.021	0.113	-0.240	-0.090	-0.088
		14-16	0.244	0.199	-0.159	-0.168	-0.042	0.001

TABLE II: Values of the Pearson correlation between the weather data and the check-ins and check-outs for two intervals of two hours in a day (from 11h to 13h and from 14h to 16h) using the data from 7/1/2019 to 28/2/2019.

B. DCCA and DCCA-l

Since the data is non-stationary DCCA was used (see Figures 5 4). According to Zebende [4], if the data only had linear trends the coefficient would tend to a constant value. Yet, as observed in Figures 5 and 5, this does not happen. In Figure 4, DCCA coefficient of check-outs, the coefficient values for temperature and humidity oscillate around zero, so there is approximately no correlation. The coefficient for the wind intensity is mostly negative and for the precipitations is mostly positive. Looking at the values in accordance with window size n, it can be seen that seasonality is most evident for n=12, n=24 and n=168, which corresponds to half a day, a day and a week respectively. In Figure 5, DCCA coefficient for check-ins against the values of temperature and wind intensity are negative, for humidity is positive and for the precipitation, it oscillates around zero. Seasonality is evident at n=24, and n=168.

Since the data is periodic and non-stationary DCCA-l was used, see Figures 7 and 6. From Figure 7, we observe that there is no correlation between the check-ins and the weather variables. From Figure 6 there is a negative correlation between the wind-intensity and the check-outs of around 0.18, and no correlation for the other variables.

VI. CONCLUSION

In this work, an exploratory analysis of the data from GIRA stations and weather was conducted. The Pearson correlation was used to study the correlations between the demand for bicycles and the weather. It was concluded that the correlation between the temperature and the demand for bicycles is positive yet weak, the correlation between the wind intensity and the demand for bicycles is negative and soft, the correlation for the precipitation and humidity are inconclusive. DCCA analysis was also used to study the correlation, confirming the non-linearity of trends and the daily-and-weekly seasonality of data, and further suggesting a negative correlation between the wind intensity and bike demand, negative correlation between check-ins and temperature and positive correlation between check-ins and humidity. DCCA-l analysis suggests that the intensity of the wind and the check-outs are correlated. As can be seen, the various methods do not seem to agree on the correlations due to the assumptions made in the use of the different statistics.

Future Work: In the future this analysis could be widen to include other groups pf stations and bigger period in time (to get a bigger variation in the weather), a study of how the aggregation of the data (used in for the pearson) changes the results could also be done. Different granularity of the data could also be tested. Taking into account the saturation of the stations, the seasonality and the spacial dependency between stations could also be incorporated in correlation analyzes to produce better results.

Acknowledgments

I want to thank Prof. Rui Henriques from INESC-ID for the support during the realization of this project and and Lisbon City Council (CML) for gently providing the data to conduct this study.

[1] Boris Podobnik and H Eugene Stanley. Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series. *Physical review letters*, 100(8):084102, 2008.

[2] C-K Peng, Sergey V Buldyrev, Shlomo Havlin, Michael Simons, H Eugene Stanley, and Ary L Goldberger. Mosaic organization of dna nucleotides. *Physical review e*, 49(2):1685, 1994.

[3] Davor Horvatic, H Eugene Stanley, and Boris Podobnik. Detrended cross-correlation analysis for non-stationary time series with periodic trends. *EPL (Europhysics Letters)*, 94(1):18007, 2011.

[4] Gillney Figueira Zebende. Dcca cross-correlation coefficient: quantifying level of cross-correlation. *Physica A: Statistical Mechanics and its Applications*, 390(4):614–618, 2011.

[5] Brian S Flynn, Greg S Dana, Justine Sears, and Lisa Aultman-Hall. Weather factor impacts on commuting to work by bicycle. *Preventive medicine*, 54(2):122–124, 2012.

[6] 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, 2011.

[7] Tom Thomas, CF Jaarsma, and SIA Tutert. Temporal variations of bicycle demand in the netherlands: The influence of weather on cycling. 2009.

[8] Christopher Gallop, Cindy Tse, and Jinhua Zhao. A seasonal autoregressive model of vancouver bicycle traffic using weather variables. In *Transportation Research Board 91st Annual Meeting*, number 12-2119, 2012.

[9] Wafic El-Assi, Mohamed Salah Mahmoud, and Khandker Nurul Habib. Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in toronto. *Transportation*, 44(3):589–613, 2017.

[10] Huthaifa I Ashqar, Mohammed Elhenawy, and Hesham A Rakha. Modeling bike counts in a bike-sharing system considering the effect of weather conditions. *Case Studies on Transport Policy*, 7(2):261–268, 2019.

Appendices

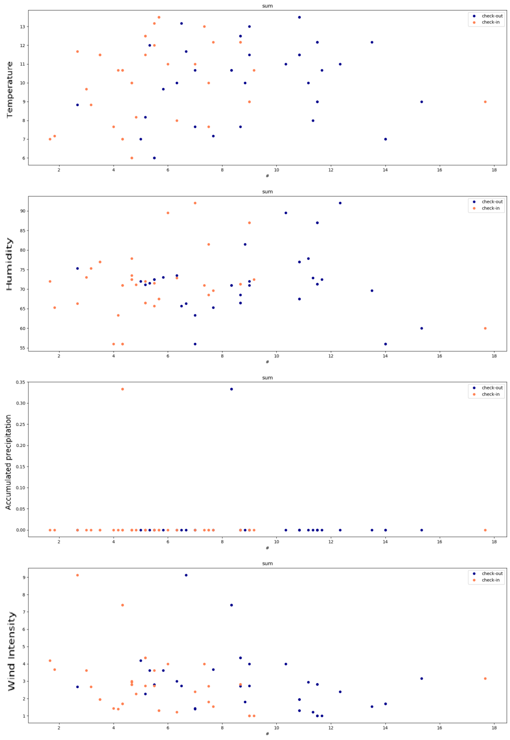

FIG. 8: Scatter plot for the sum of check in and outs and the weather data for 14h to 16h

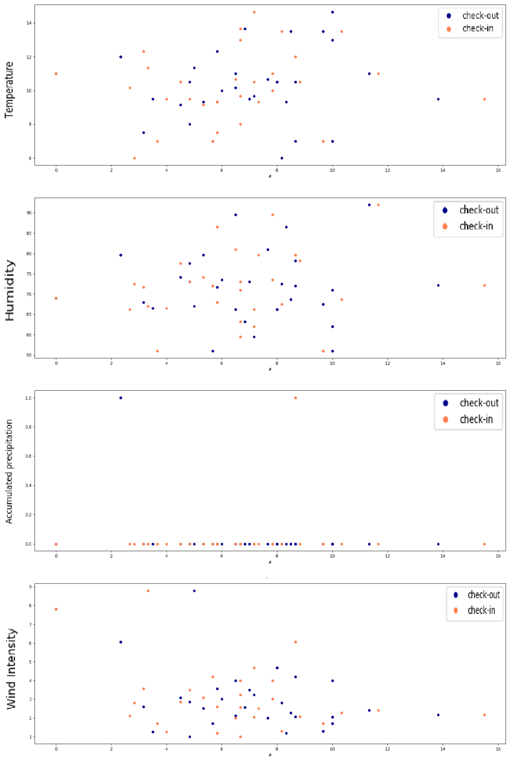

FIG. 9: Scatter plot for the sum of check in and outs and the weather data for 11h to 13h

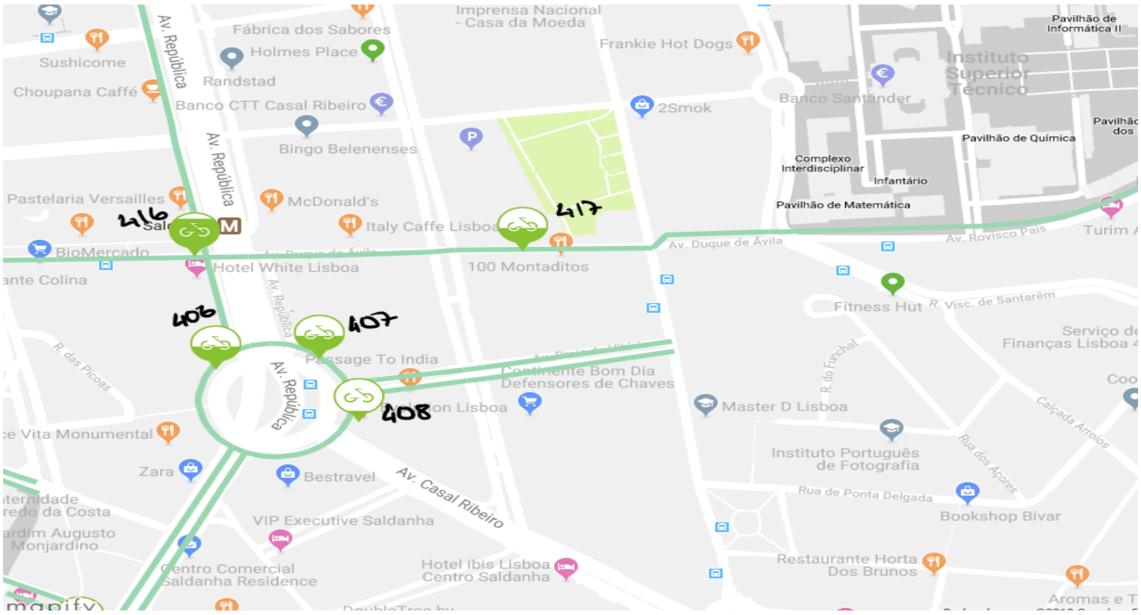

FIG. 10: Map with location of the GIRA stations

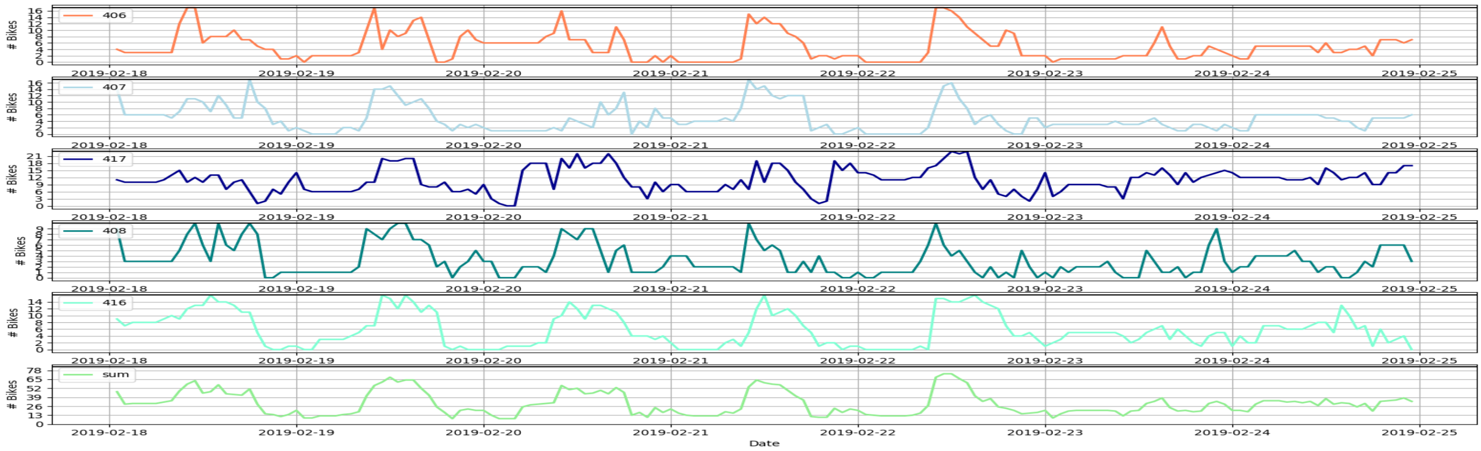

FIG. 11: Number of bicycles in the various stations and the sum of the bicycles parked in those station from 18/2/2019 and 24/2/2019

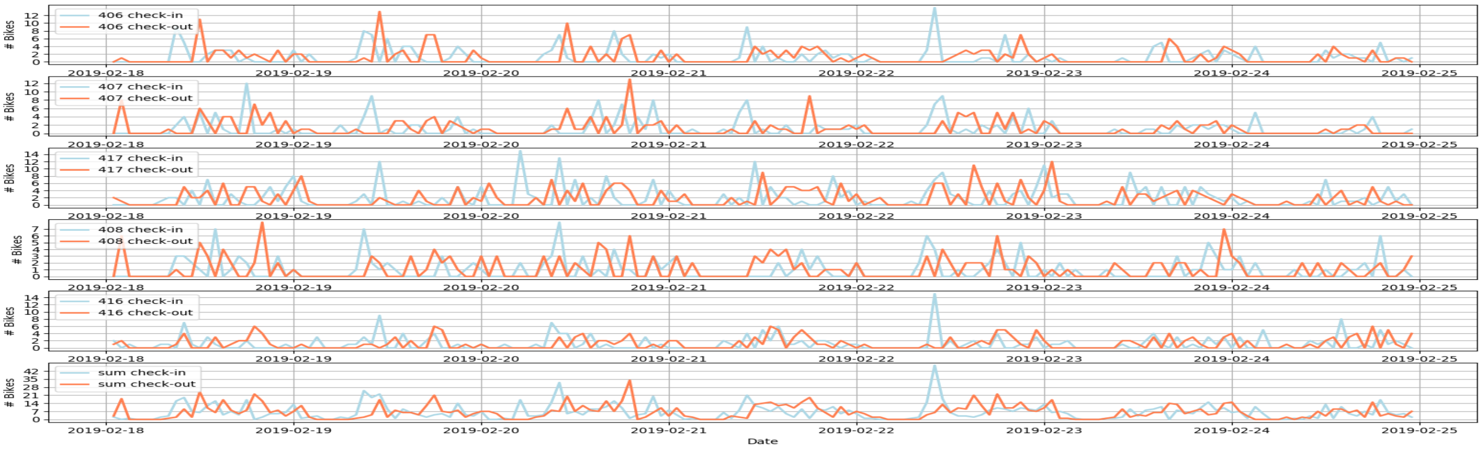

FIG. 12: Number of bicycles that were removed or parked in the various stations and the sum of the bicycles parked in those station from 18/2/2019 and 24/2/2019