Study of Correlation between the weather and Bicycle demand

INTRODUCTION TO RESEARCH 2019/2020

Ana Santos 84364

Correlation between bike demand and weather

Index

- Introduction to the problem
- Data Preparation
- Background and Methodology
- Results and discussion
- Conclusions
- Future Work

Introduction - ILU

ILU address two major **challenges**:

1. Lack of **integrative analysis** to **combine different sources of urban data** collected from public transport
2. Absence of **situational context** in **predictions** and **public transport planning**

Introduction – Objectives

This work focuses on the **bicycle mobility** in particular **the public sharing system** in Lisbon (GIRA).

The main objectives are:

1. Exploratory analysis of GIRA's data
2. **Consolidate** the data from **GIRA** with data from **weather** stations
3. Study the **correlation** between the data sources in order to understand the **effect of weather on bicycle demand.**

Data Preparation – Bike sharing system

Bikes can be removed form a dock (**check-in**) and returned to another dock (**check-out**).

The number of check-ins and check-outs give a **measure of bicycle demand** 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)

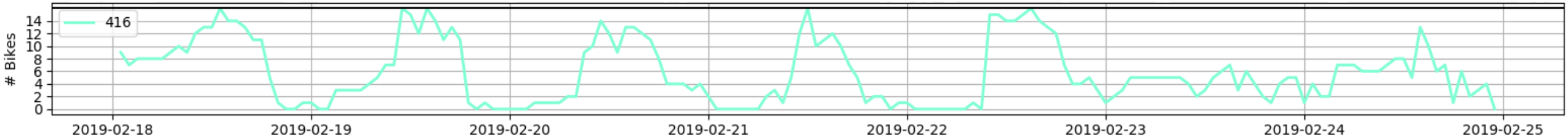


FIG1 : Number of bicycles in station 416 from GIRA from 18/2/2019 to 21/2/2019

Data Preparation – Bike sharing system

To prevent this masking of the demand **5 stations were grouped**.

The **sum** of the number of bikes, check-ins and check-outs in the group was considered.

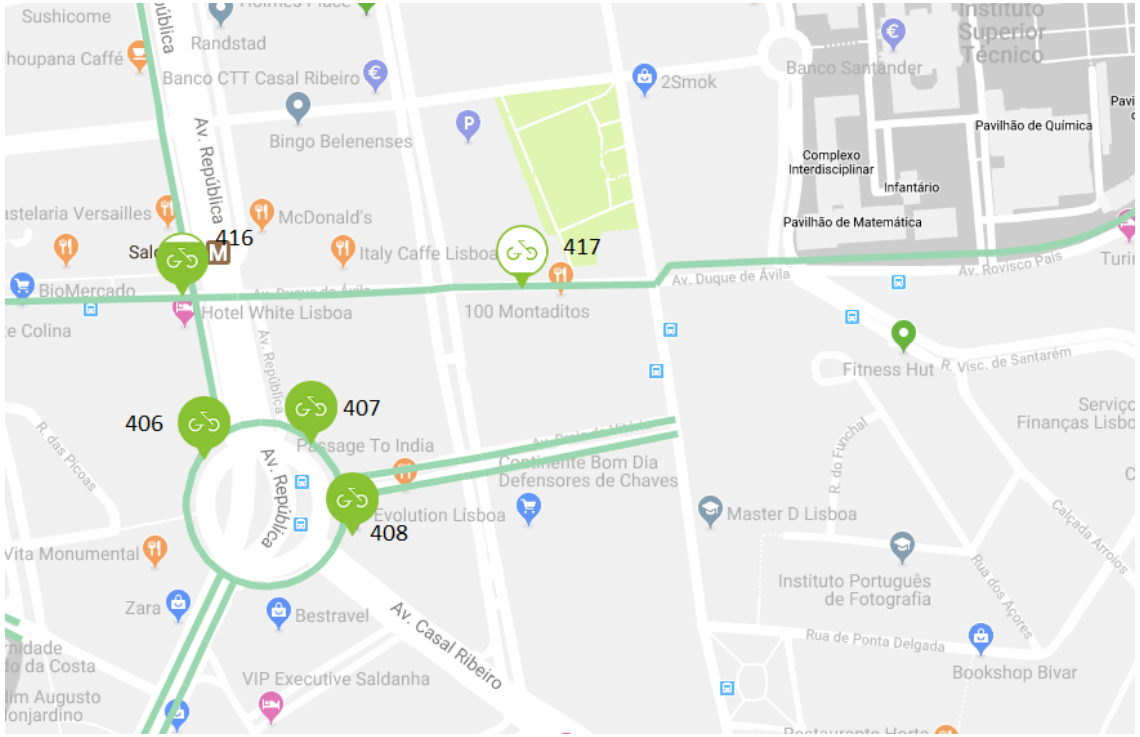


FIG2 : Map of the group of stations from the GIRA network

Data Preparation– Bike sharing system

Time series were created with granularity of **60 minutes**, for the **number of bikes**, **check-ins** and **check-outs**.

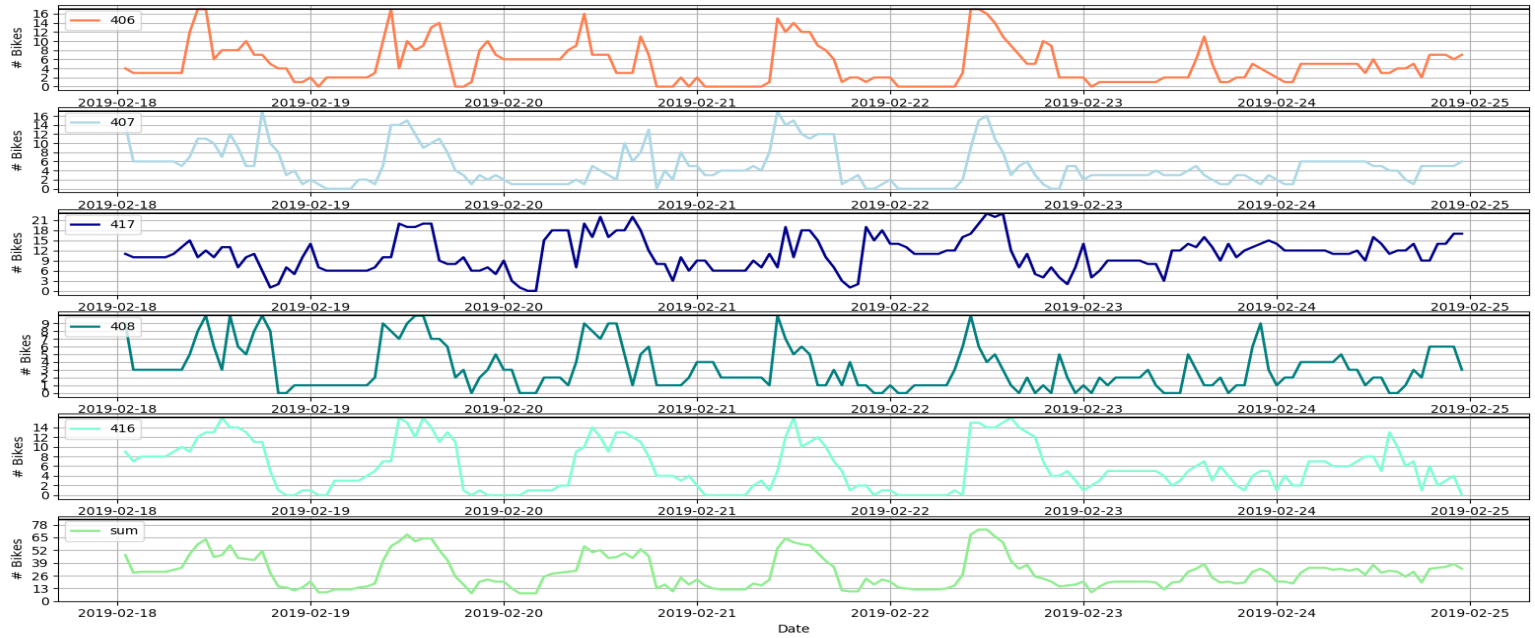


FIG3 : Number of bicycles for the stations 406, 407, 408, 416, 417 and the sum from GIRA from 18/2/2019 to 21/2/2019

FIG4 : Number of check-ins (in blue) and check-outs (in orange) for the stations 406, 407, 408, 416, 417 and the sum from GIRA from 18/2/2019 to 21/2/2019

Data Preparation - Weather

The weather stations collect/capture data on **temperature** (ºC), **humidity** (%), **wind intensity** (km/h) and **accumulated precipitation** (mm).

The station chosen was 579.

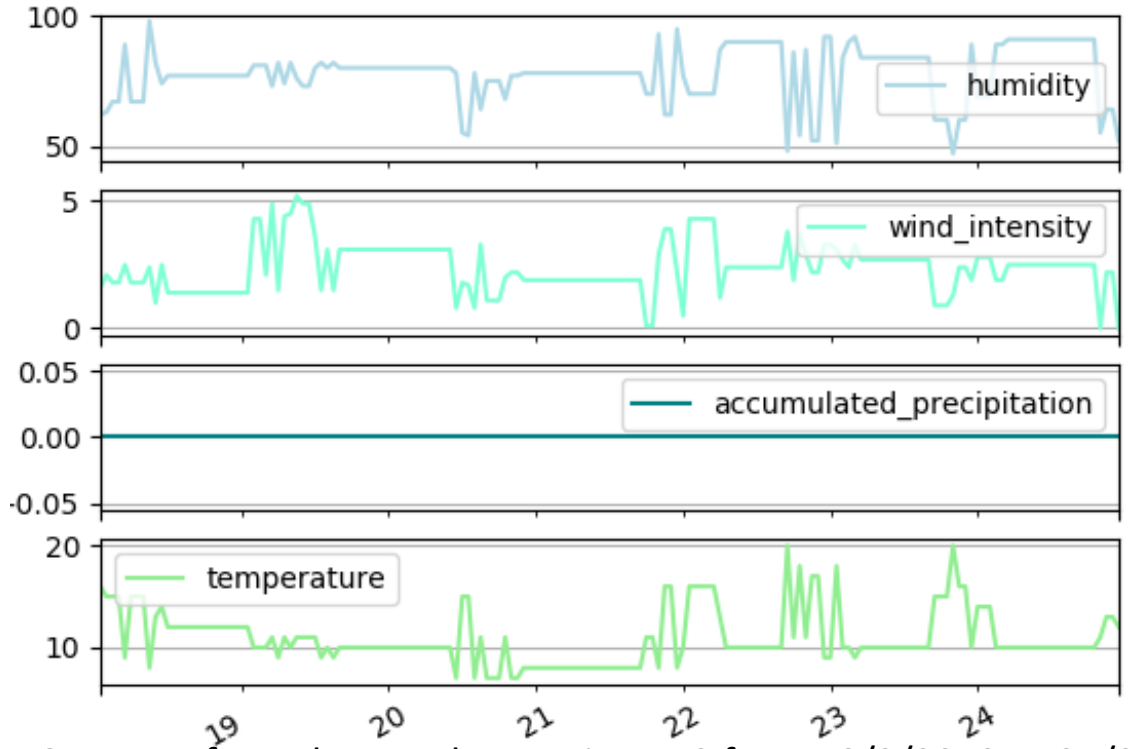


FIG5 : Data from the weather station 579 from 18/2/2019 to 21/2/2019

Background and Methodology – Pearson Correlation

The Pearson correlation coefficient (**PCC**) is a measure of the **linear correlation** between two **time series**, x and y, given by $r_{x,y}$.

PCC's of values **-1** or **1** imply that a **linear equation** describes the relationship between x and y perfectly.

PCC's value of **0** implies that there is **no linear correlation.**

To remove the correlation due to the daily and weekly periodicity, for every working day the average of two hour off peak hours, was used.

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

Background and Methodology – DCCA

Boris Podobnik [1] presents an alternative method, DCCA is used to investigate **power-law cross-correlations** between time series in the presence of **non-stationarity**. If we have time series x and x':

1. First we start by calculating the **integrated signals** : $R_k = \Sigma_{i=1}^{k} x_i$
2. Time series is divided into T-n **overlapping boxes**.
3. For each box (i), a **least square regression** is made: $\widetilde{R_{k,i}}$
4. The **detrended walk** is calculated: $(R_k - R_k{}')$
5. The covariance of the residuals in each box is calculated: $f_{DCCA}^2(n,i) = \frac{1}{n-1}\Sigma_{k=i}^{i+n}(R_k - \widetilde{R_{k,i}})(R_k{}' - \widetilde{R_{k,i}{}'})$
6. The detrended covariance is then calculated: $F_{DCCA}^2(n) = \frac{1}{N-n}\Sigma_{i=1}^{T-n} f_{DCCA}^2(n,i)$

If the signal are the same the detrended covariance reduces to the detrended variance F_{DVA}^2 used in the detrended fluctuation analysis (DFA) method [2].

Background and Methodology – DCCA-*l*

Horvatic et al [3] proposed a new method DCCA-l for time series with periodic trends. If we have time series x and x':

1. First we start by calculating the **integrated signals** : $R_k = \Sigma_{i=1}^{k} x_i$
2. Time series is divided into T-n **overlapping boxes**.
3. For each box (i), the signal is fitted to a polynomial of order l, that varies with the size of the box n, is made: $\widetilde{R_{k,i}}$
4. The **detrended walk** is calculated: $(R_k - R_k{}')$
5. The covariance of the residuals in each box is calculated: $f_{DCCA}^2(n,i) = \frac{1}{n-1}\Sigma_{k=i}^{i+n}(R_k - \widetilde{R_{k,i}})(R_k{}' - \widetilde{R_{k,i}{}'})$
6. The detrended covariance is then calculated: $F_{DCCA}^2(n)\alpha\frac{1}{N-n}\Sigma_{i=1}^{T-n} f_{DCCA}^2(n,i)$

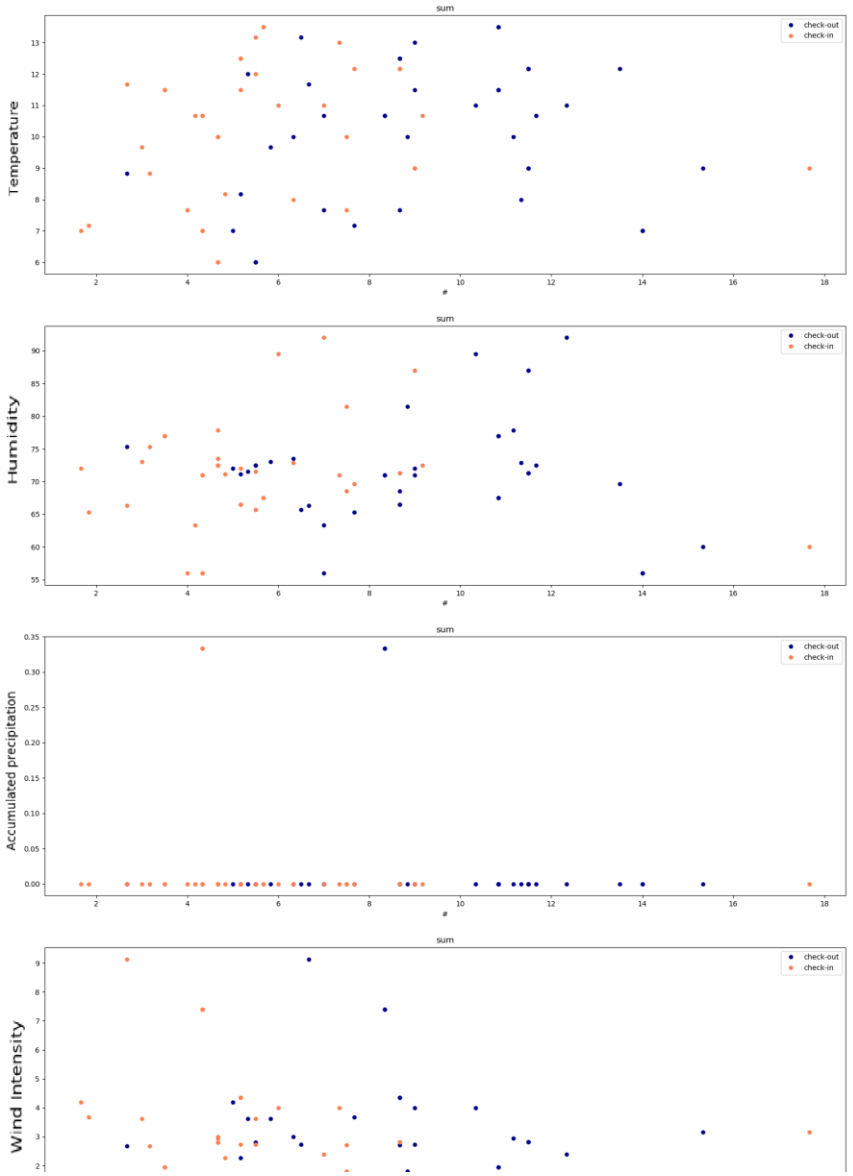
Background and Methodology – DCCA-I coefficient

The introduces methods do not properly quantify the level of cross-correlation.

So a coefficient is proposed based on the DCCA methods.

This coefficient σ has values between -1 and 1.

$$\sigma_{DCCA} = \frac{F^2_{DCCA}\{x,x\prime\}}{F_{DFA}\{x\}F_{DFA}\{x\prime\}}$$



Results and discussion – Pearson correlation

The scatter plots suggest a lack of linear correlation.

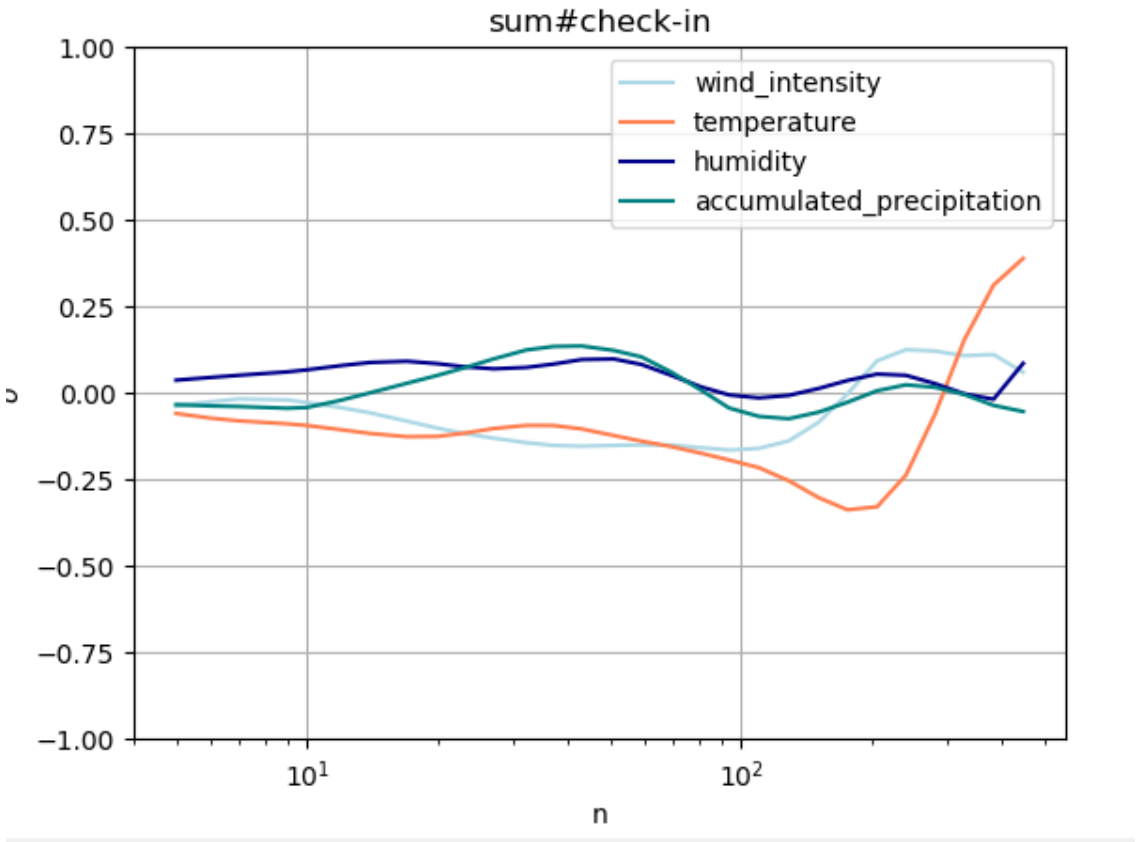
FIG6 : Scatter plot for the sum of check-ins (orange) and check-out(blue) and the weather for 14-16 h from 7/1/2019 to 28/7/2019

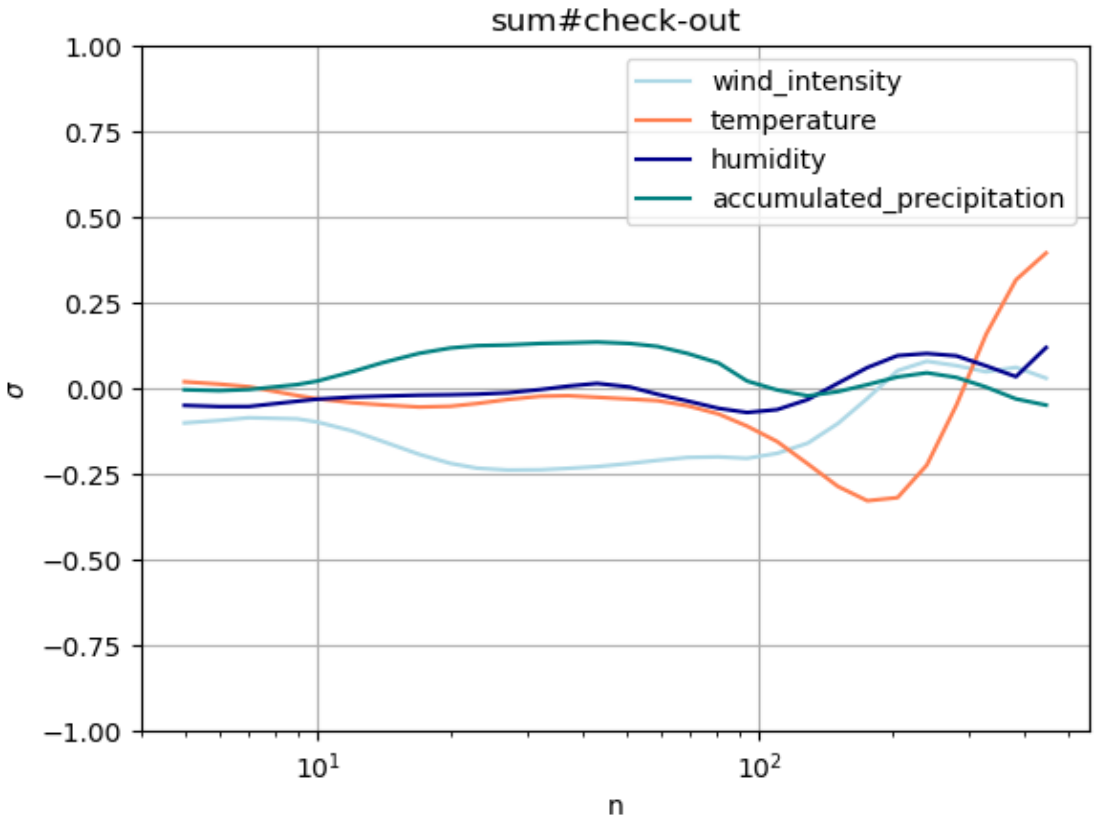
station			406	407	408	416	417	sum
temp	check-in	11-13	0.147	0.178	0.491	0.043	0.050	0.239
		14-16	0.127	0.255	0.050	0.050	0.088	0.138
	check-out	11-13	0.112	-0.171	0.273	0.190	-0.057	0.090
		14-16	0.303	0.082	-0.065	-0.065	0.115	0.167
prec	check-in	11-13	0.124	0.161	0.151	0.251	-0.070	0.161
		14-16	-0.204	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

TABI : 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.

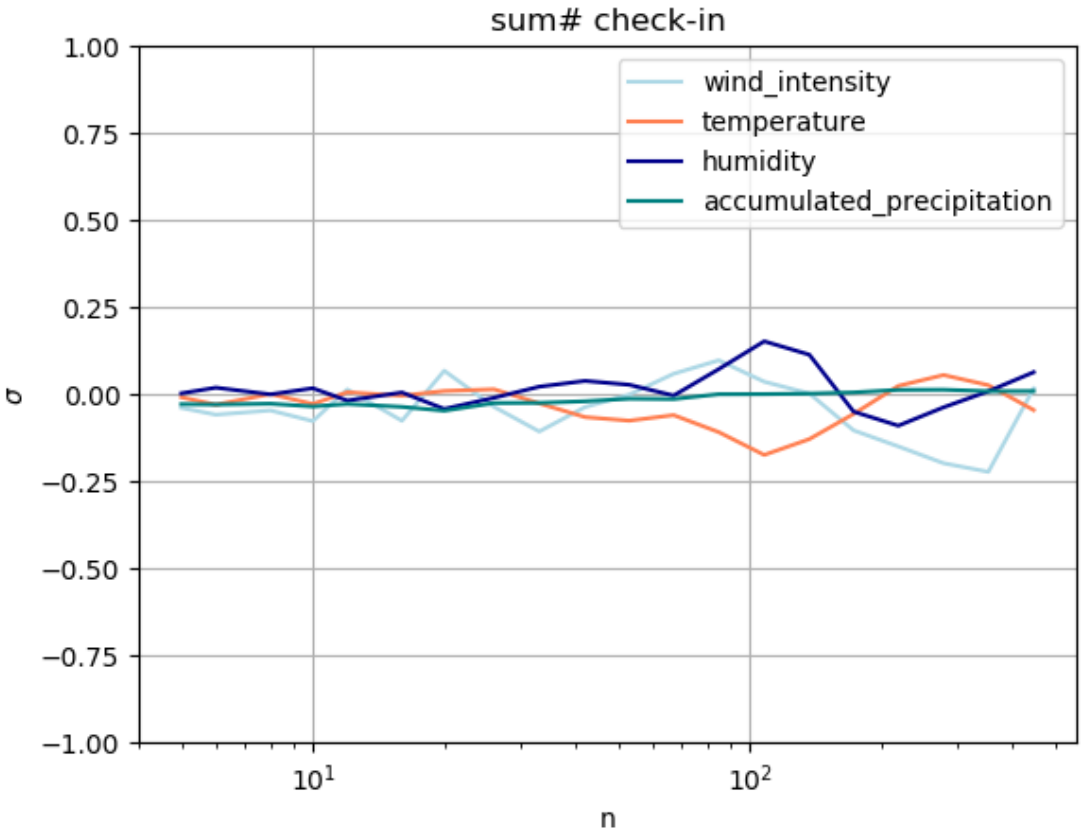
Results and discussion – Pearson Correlation

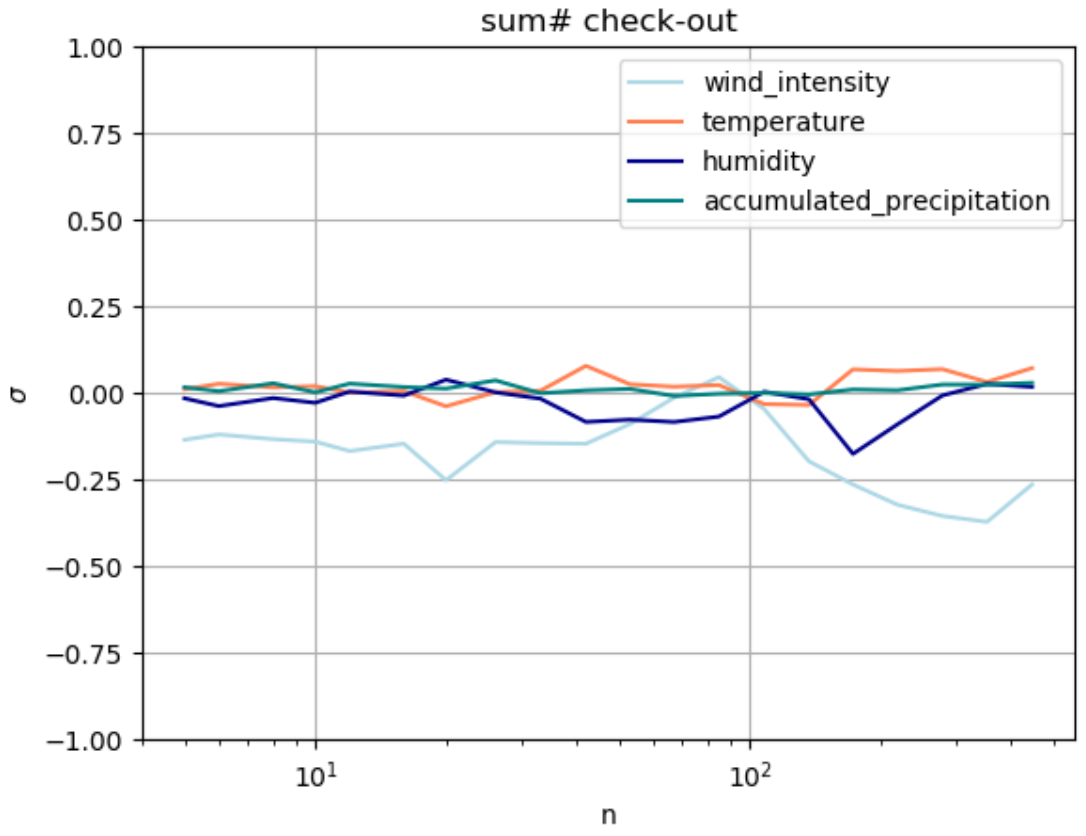
Results and Discussion - DCCA





Results and Discussion – DCCA-/





Conclusions

- 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. 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 **nonlinearity** 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**.
- DCCA-l analysis suggests that the **intensity of the wind** and the **check-outs** are correlated.

Future Work

- In the future this analysis could be widen to **include other groups of stations** and **bigger period in time**.
- 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 **seasonality** and **the spatial dependency between stations** could also be incorporated in correlation analyzes to produce better results.

References

[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.

Thank you