Capgemini test: forecasting water levels

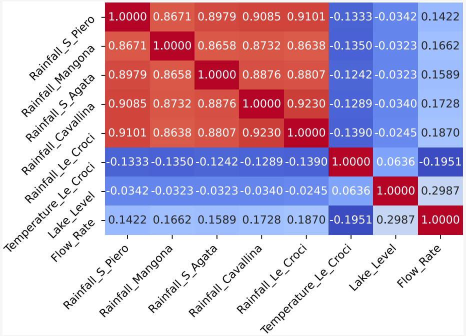
Stefano Petrucci
petrucci.ste@gmail.com

Software

- Written in Python and available on GitHub
- Used standard libraries:
  - Pandas
  - Seaborn
  - Statsmodels
- One class for the data set (with methods for plotting and split data into train/test)
- Some functions for the statistical models
- Everything is documented within the code (it should be Doxigen friendly as well)

Data set

- Chosen *Lake_Bilancino*
- Two target variables:
  - Lake level
  - Flow rate
- Most variables missing before 01/01/2004
  $\Rightarrow$ removed ($\sim$ 8% of the total)
- Missing data not replaced[a]



	Rainfall_S_Piero	Rainfall_Mangona	Rainfall_S_Agata	Rainfall_Cavallina	Rainfall_Le_Croci	Temperature_Le_Croci	Lake_Level	Flow_Rate
Rainfall_S_Piero	1.0000	0.8671	0.8979	0.9085	0.9101	-0.1333	-0.0342	0.1422
Rainfall_Mangona	0.8671	1.0000	0.8658	0.8732	0.8638	-0.1350	-0.0323	0.1662
Rainfall_S_Agata	0.8979	0.8658	1.0000	0.8876	0.8807	-0.1242	-0.0323	0.1589
Rainfall_Cavallina	0.9085	0.8732	0.8876	1.0000	0.9230	-0.1289	-0.0340	0.1728
Rainfall_Le_Croci	0.9101	0.8638	0.8807	0.9230	1.0000	-0.1390	-0.0245	0.1870
Temperature_Le_Croci	-0.1333	-0.1350	-0.1242	-0.1289	-0.1390	1.0000	0.0636	-0.1951
Lake_Level	-0.0342	-0.0323	-0.0323	-0.0340	-0.0245	0.0636	1.0000	0.2987
Flow_Rate	0.1422	0.1662	0.1589	0.1728	0.1870	-0.1951	0.2987	1.0000

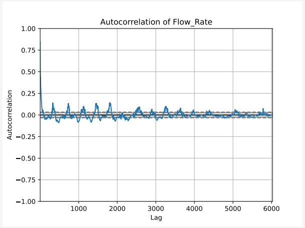
Correlation matrix

[a] Applied an interpolation only when computing the autocorrelation.

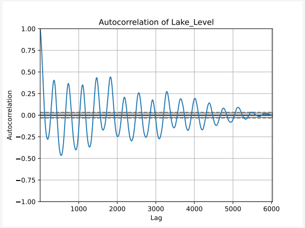
Forecasting strategy

1. Simple AutoRegressive (AR) model
2. More complex AutoRegressive Integrated Moving Average (ARIMA) model
3. Multivariate analysis (not implemented)

Both models used in this project require to setup the lag.
⇒ chosen from autocorrelation plots.



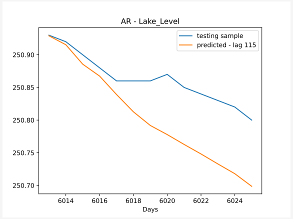
Flow rate - autocorrelation



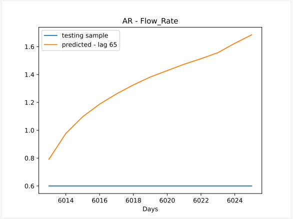
Lake level - autocorrelation

Predictions - 13 samples

AutoRegressive model

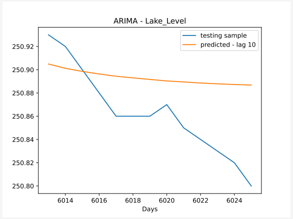


Lake level

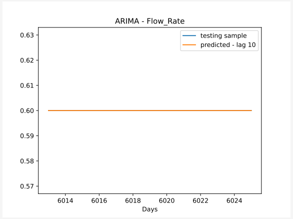


Flow rate

ARIMA model



Lake level



Flow rate

Conclusions

- Implemented a toy script to compute AR and ARIMA algorithms
- Both algorithms showed better performance on `Lake_level`
- The behavior of `Flow_rate` requires additional investigation
  Optimizing ARIMA's parameters might improve performance

Thank you for your attention