Brief Scientific Report on: Air Quality forecasting

By : Osias Noël Nicodème Finagnon TOSSOU

Project on GitHub : https://github.com/osiastossou/airqo-test.git

I- Brief Explanation of the dataset

Air is one of the elements that allow man to survive. Thus, it is important that its quality be monitored to avoid diseases.

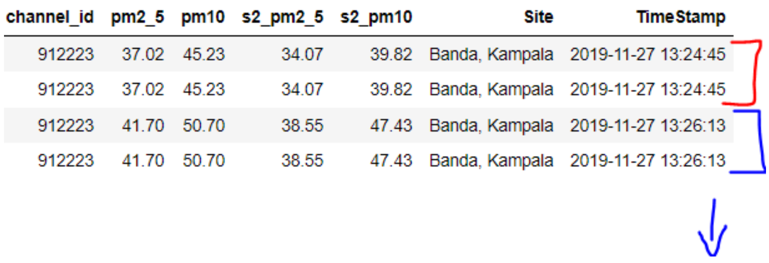
This report presents the data modeling work to predict air quality over time.

We received a dataset of about 2 GB. This dataset contains property records measuring air quality at 42 sites in 3 years (2019 - 2020 - 2021).

In the paper we will present an analysis of the data and modeling for the prediction of air quality property values each hour of the day.

II- Data preprocessing

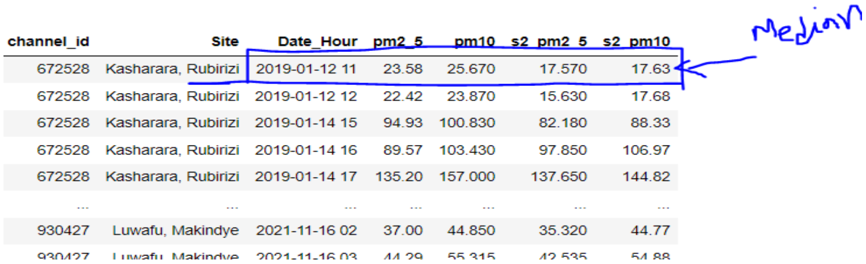
- In the dataset, we have noticed that there are duplicate data



channel_id	pm2_5	pm10	s2_pm2_5	s2_pm10	Site	TimeStamp
912223	37.02	45.23	34.07	39.82	Banda, Kampala	2019-11-27 13:24:45
912223	37.02	45.23	34.07	39.82	Banda, Kampala	2019-11-27 13:24:45
912223	41.70	50.70	38.55	47.43	Banda, Kampala	2019-11-27 13:26:13
912223	41.70	50.70	38.55	47.43	Banda, Kampala	2019-11-27 13:26:13

Delete the duplicate row in data.

- In the dataset, the data are recorded several times in the interval of an hour (Ex: 13:24:45 and 13:26:13 in the previous image), we have taken the median of the values of the properties we have replaced for a given hour of a day.

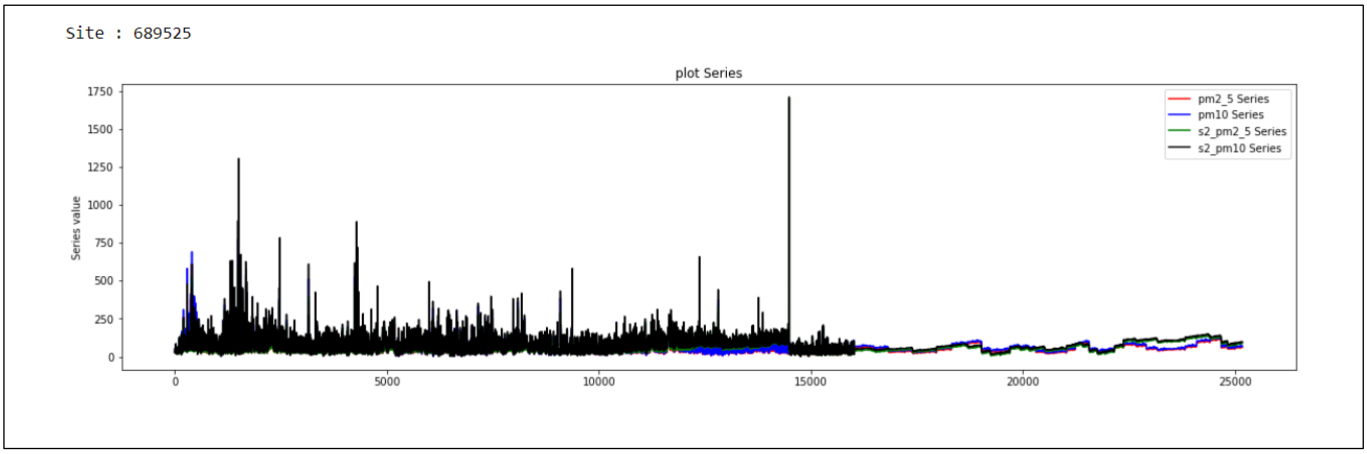


Median

channel_id	Site	Date_Hour	pm2_5	pm10	s2_pm2_5	s2_pm10
672528	Kasharara, Rubirizi	2019-01-12 11	23.58	25.670	17.570	17.63
672528	Kasharara, Rubirizi	2019-01-12 12	22.42	23.870	15.630	17.68
672528	Kasharara, Rubirizi	2019-01-14 15	94.93	100.830	82.180	88.33
672528	Kasharara, Rubirizi	2019-01-14 16	89.57	103.430	97.850	106.97
672528	Kasharara, Rubirizi	2019-01-14 17	135.20	157.000	137.650	144.82
...	...	...	...	...	...	...
930427	Luwafu, Makindye	2021-11-16 02	37.00	44.850	35.320	44.77
930427	Luwafu, Makindye	2021-11-16 03	44.29	55.315	42.535	54.88

- In the dataset, we noticed that there are missing data for some hours of some days. So we decided to do an imputation with a regression method by using the site metadata.

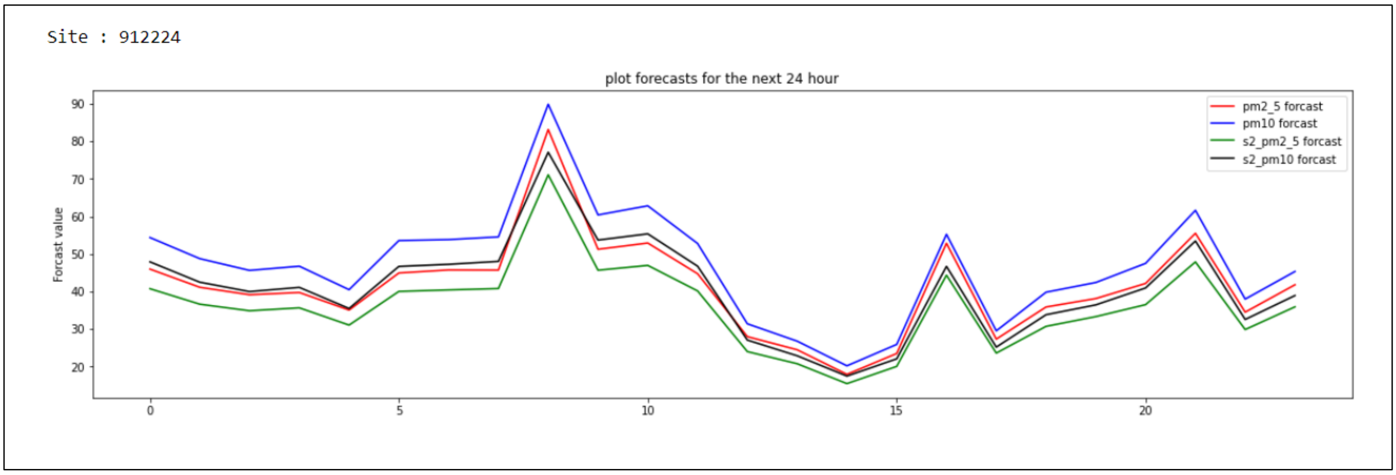
- We visualize all the series data, as we can see here for one site. (You can see in detail in the Jupyter NoteBook[i])



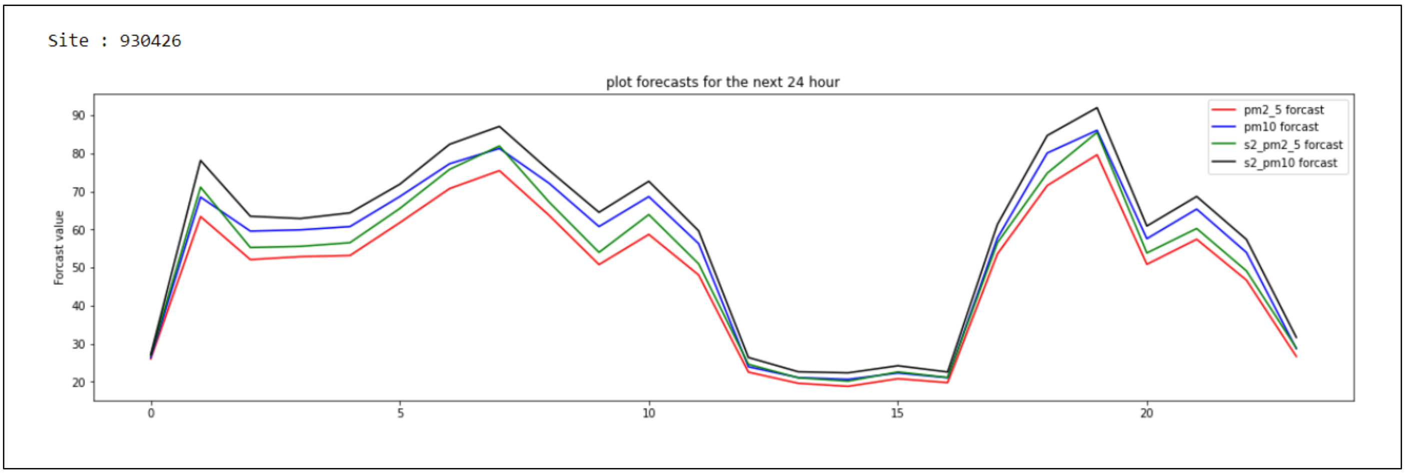
```
Site : 689525
```

III- Modeling and forecasting

The data we have are following a time so we have a problem of modeling timeseries with periodicities. There are several models of timeseries, here we used the SARIMA model and make an evaluation with the RMSE by site. (See the Jupyter NoteBook for the detail)



```
Site : 912224
```



```
Site : 930426
```

IV- Summary

We proposed a forcasting model of the values of each property qualifying the nature of the air. It should be noted that a preprocessing work has been done on the data (Removal of duplicate data, Imputation of missing data by a Machine Learning model)

[i] https://github.com/osiastossou/airqo-test.git