Statistical Learning Project

2st Milestone

G08

Prediction and analysis of air quality

Riccardo Ceccaroni, Giusy Beatrice Colarusso, Gabriele Giannotta, Francesco Lauro

Abstract

Air pollution is one of the big issues that affect metropolitan areas all around the world. Traffic and factories play a significant role. We need to implement models that keep track of data about the concentration of air pollutants (SO_2, NO_2, etc.). Indeed, the diffusion of these toxic gases in the air affects people's quality of life. The increase in data availability and new discoveries in terms of computational technologies are making it possible to predict and analyse the air quality, providing extremely useful information to control air pollution.

Main research aim & framework

The main goal of the project is to find a model that manages to predict air quality. Our interest in this topic stems from the fact that we think it's important to know more about air quality situation. Indeed, we strongly believe that predictions could help reducing pollution. Living in a "unique" historical period, namely the Covid-19 pandemic, a secondary goal of our research is understanding if this virus, besides changing our habits, has had an effect on air quality ("Air Pollution Goes down as Europe Takes Hard Measures to Combat Coronavirus" 2020).

Data collection & source(s)

Data will be collected on a daily basis (about 1 record per hour) from the Air Quality Programmatic APIs website. They will be collected into a dataframe through our own Python script. This script connects directly to the website using APIs, and requests the air quality data measured in the following data-gathering entity in the city of Rome: Fiumicino Guglielmi, via Francia, Cinecitta, Villa Ada, Cipro, via Arenula. Each record will be used to fill in a dataframe which will be saved into a .csv format file. Finally, all these dataframes will be merged together.

From each data-gathering station we will collect the following data: date, time, location, carbon monoxide CO, nitrogen dioxide NO_2, ozone O_3, sulphur dioxide SO_2, particulate matter PM_{10}, particulate matter $PM_{2.5}$, humidity h, pressure p, temperature t, and wind speed w.

The main concern in recording data is that data-gathering entities are uploaded at a different time during the day, and some of them are updated only once a day. For this reason, we could be having some stations with more recordings than others.

index	Local Date	Local Time	Latitudine	Longitudine	BENZENE	CO	NO	NO2
1	2020-03-27	12:00	41.90636	12.44760	NA	NA	4	15
2	2020-03-27	12:00	41.89402	12.47537	NA	NA	0	11
3	2020-03-27	13:00	41.76819	12.23705	NA	NA	1	7
4	2020-03-27	13:00	41.94745	12.46959	0.7	NA	8	14
5	2020-03-27	13:00	41.85772	12.56866	NA	NA	0	10

index	NOX	O3	PM10	PM2.5	SO2	temperature	precipitation	humidity	pressure
1	21	56	6	2	NA	16.0	NA	59	1008
2	11	60	9	2	NA	16.0	NA	59	1008
3	9	75	6	3	NA	14.2	0.2	65	1009
4	27	NA	7	3	NA	16.0	NA	63	1009
5	10	68	9	4	NA	16.0	NA	63	1009

Model & Methods update

We are going to divide the data we gathered in two phases. The first one contains observations starting from March the 24th to May the 4th, while the second one starts from May the 5th until the day we stop collecting data. We will leave out a few days of the first phase and use the remaining ones to approximate the posterior distribution through the **INLA** technique (Krainski et al. 2018). Then we will check the quality of the prediction using the days we didn't include in the previous fitting process. Afterward, will make predictions on the second phase's days. Finally, we are going to compare these projections with the actual data we have for the second phase.

Software/Hardware Toolkit update

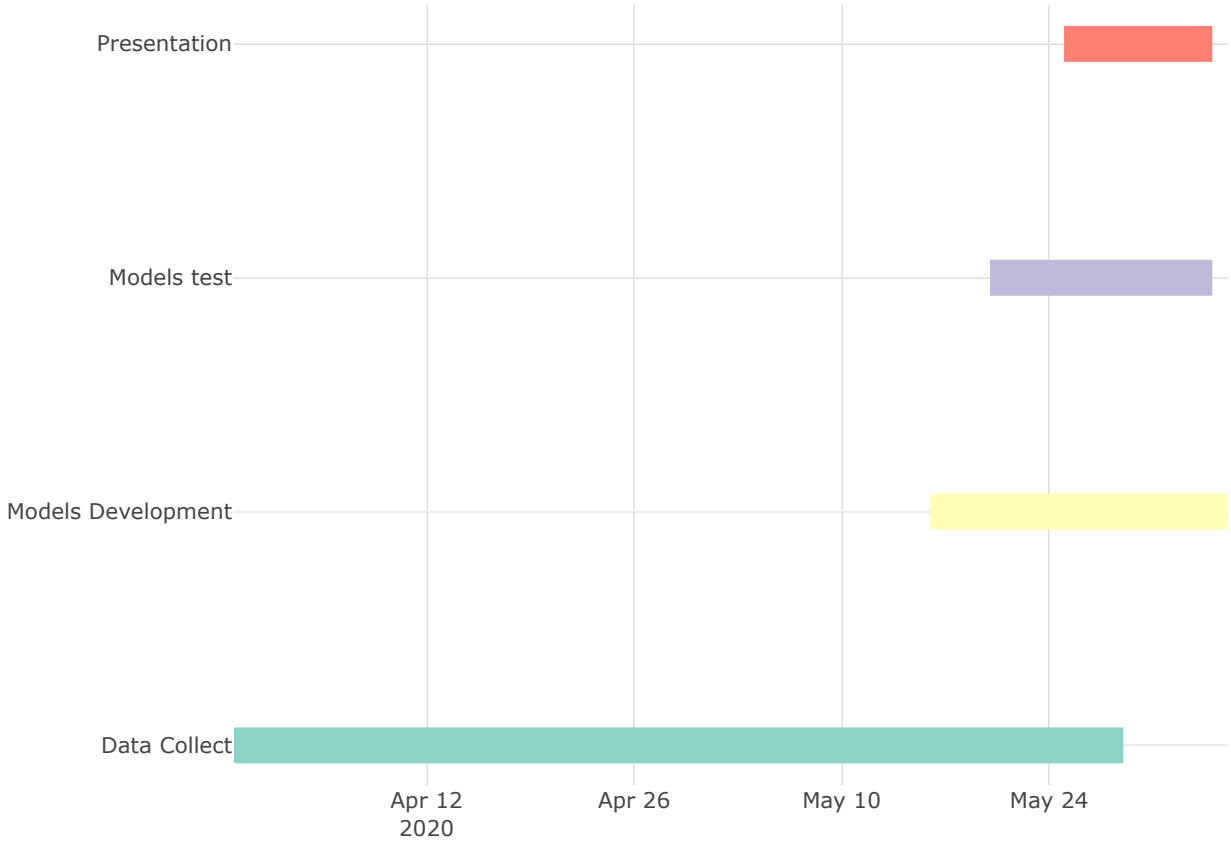
Up to this point, the only programming language we have been using is *Python*. Our Python script is automatically runned online through pythonanywhere website.

It will be the main language we are going to use for the entire data analysis, together with *R*. Moreover, we will employ the INLA library.

Problems so far...

Up to now, there have been essentially two problems: missing or incomplete data for the variables considered, even though this is not a big issue since the INLA technique can deal with missing data; doubts regarding how to handle the INLA technique.

Project Timeline update



References update

"Air Pollution Goes down as Europe Takes Hard Measures to Combat Coronavirus." 2020. https://www.eea.europa.eu/highlights/air-pollution-goes-down-as.

Cameletti, Michela, Virgilio Gómez Rubio, and Marta Blangiardo. 2019. "Bayesian Modeling for Spatially Misaligned Health and Air Pollution Data Through the Inla-Spde Approach." *Spatial Statistics* 31 (April). https://doi.org/10.1016/j.spasta.2019.04.001.

Delavar, Mahmoud, Amin Gholami, Gholam Shiran, Yousef Rashidi, Gholam Nakhaeizadeh, Kurt Fedra, and Smaeil Afshar. 2019. "A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran." *ISPRS International Journal of Geo-Information* 8 (February): 99. https://doi.org/10.3390/ijgi8020099.

Kleine Deters, Jan, Rasa Zalakeviciute, Mario Gonzalez, and Yves Rybarczyk. 2017. "Modeling Pm 2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters." *Journal of Electrical and Computer Engineering* 2017 (June): 1–14. https://doi.org/10.1155/2017/5106045.

Krainski, Elias, Virgilio Gómez Rubio, Haakon Bakka, A. Lenzi, Daniela Castro-Camilo, Daniel Simpson, Finn Lindgren, and Håvard Rue. 2018. *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and Inla*. https://doi.org/10.1201/9780429031892.

Qi, Zhongang, Tianchun Wang, Guojie Song, Weisong Hu, Xi Li, Zhongfei, and Zhang. 2017. "Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-Grained Air Quality." *IEEE Transactions on Knowledge and Data Engineering* PP (November). https://doi.org/10.1109/TKDE.2018.2823740.

Venegas, Laura, Nicolás Mazzeo, and Mariana Dezzutti. 2014. "A Simple Model for Calculating Air Pollution Within Street Canyons." *Atmospheric Environment* 87 (April): 77–86. https://doi.org/10.1016/j.atmosenv.2014.01.005.

Wang, Ping, Yong Liu, Zuodong Qin, and Guisheng Zhang. 2014. "A Novel Hybrid Forecasting Model for Pm10 and So2 Daily Concentrations." *The Science of the Total Environment* 505C (November): 1202–12. https://doi.org/10.1016/j.scitotenv.2014.10.078.