Predictive air pollution and environmental variables in Covid-19

Gómez A. and Torres A.

$31^{\underline{st}}$ of March, 2021
DATATHON - V Bioinformatics Workshop UGR

I. The data and chosen variables

After a bibliographic research of the influence of the transmission and mortality of COVID-19, it is suggested that the air pollution and environmental factors could play a key role. In fact, air pollution seems to encourage the virus transmission. Whereas the environmental factors could have a negative correlation according to some bibliographic references ([1]–[4]). Although the causes of these relationships are still not clear.

To carry out our study, we loaded the data from Github's DatA web-server. We considered the following predictive variables: **CO**, **NO2**, **O3**, **SO2**, **PM10** and **PM2.5** as the *air pollution variables* and **Temperature**, **Windspeed**, **Precipitation** and **Insolation** as the *environmental variables*. The *response variables* are **newCasesTotal**, **newDeaths**, **newHospitalized**, **newUCI** and **pDeaths.cases**.

II. Descriptive analysis and rates of interest

We did a descriptive analysis of the new daily COVID-19 cases, as well as some **clustering** by provinces and communities. We drew a dendrogram according to the new daily cases, in order to **detect patterns**. For instance, the influence of the proximity between populations or population density.

Moreover, we tried to understand the statistical properties of the time series and analyzed the data as a time series. For that, we tested the **existence of an auto-correlation of the time series in the new daily cases**. We did a prediction in the cases data with an **ARIMA model**, which **detected a pattern of peaks up and down** on the new daily cases data frame.

Finally, we calculated the **cumulative incidence** using the median every two weeks. Thus, we could compare this rate in different regions, such as Catalonia, Madrid or Andalusia.

III. The influence of the chosen variables in our data

We studied the data by provinces, creating only one final data frame with the median values of the same variables in the different provinces grouped by weeks. The rows of the data frame are the weeks and the columns are the variables (see our *PowerPoint* for further insight). Then we applied five lineal models (one per response variable) with the data of this data frame.

To achieve the above, we did first the **preprocessing** of the data:

- The first four response variables already suggested should be normalized:

$$\frac{Y.va}{population} \times 100000$$

  Y.va is the variable of interest -it is a data frame- and *population* is the population of each province. We should not forget to multiply by 100000 inhabitants.
- We removed the provinces from the data frames that only have *NA* values and multiplied the data frame CO by 1000 so that all the data are on the same scale ($\mu g/m^3$). Then, we kept only the common provinces in all our variables.
- Since we were in lockdown, we kept the data from **06.21.2020 to 02.22.2021**, because the fact of not using the cars or leaving the house could bias the study.
- We removed the columns that have more than 20 NAs and then did another filtering to keep only those columns (remember, they are the provinces) that are coincident.
- Create the structure of the data frame by province, where **each column will be a variable, and the rows are the different dates**. Then, we created different data frames per province (17 in total).
- We created a mega data frame joining all the created data frames (all of them have the same structure) by rows.

Finally, for the **models** we **removed the influential points** and **grouped** the data **by weeks** summarized by the **median**, obtaining a final data frame with only 34 rows (weeks). Thus, we could apply the *step* function and obtain the best general models per response variable.

IV. Results and discussion

We found that **CO** and **wind speed** have a **positive correlation on the virus spread**, whereas for the **O3**, it is negative. These results accord with the bibliography ([1]–[4]).

Regarding **NO2**, **pm2.5** and **insolation**, the results are not conclusive. Nonetheless, we suggest that **pm2.5** could have a positive influence on the severity of the disease.

Surprisingly, we did not obtain that the *temperature* has any strong correlation, instead we get that it is the *insolation*. It can be explained by a higher correlation value of the insolation with regard the temperature, as shown in the *corrplot* in the *PowerPoint*.

Table 1 Main results from the five models per response variable using *step* function. The rows are the predictive variables and the columns are the response variables. Thus, each column corresponds to a different model, being the last row the adjusted R^2 of the model. In green, the selected variables with *step*. In blue, the variables with a level of significance greater than 0.05. The signs "+" or "-" denote whether the coefficient is positive or negative and the simbols ".", "*", "**" and "***" are the level of significance.

	Cases	Hospitalizations	ICU	Deaths	pDeaths/cases
pm2.5	- *	- .	- *	-	+ **
pm10		+ .	+ .	+ .	- .
co	+ .	+ ***	+ *	+ **	- .
no2	+ *				- ***
o3		- **	- *	- ***	
temperature			- .		
wind speed		+ *		+ ***	
precipitation		-			
insolation					+ **
R^2	0.49	0.62	0.57	0.64	0.85

REFERENCES

[1] M. Ahmadi, A. Sharifi, S. Dorosti, S. Jafarzadeh Ghoushchi, and N. Ghanbari. Investigation of effective climatology parameters on covid-19 outbreak in iran. *Science of The Total Environment*, 729:138705, 2020.

[2] C. Copat, A. Cristaldi, M. Fiore, A. Grasso, P. Zuccarello, S. S. Signorelli, G. O. Conti, and M. Ferrante. The role of air pollution (pm and no2) in covid-19 spread and lethality: A systematic review. *Environmental Research*, 191:110129, 2020.

[3] M. H. Shakil, Z. H. Munim, M. Tasnia, and S. Sarowar. Covid-19 and the environment: A critical review and research agenda. *Science of The Total Environment*, 745:141022, 2020.

[4] A. Srivastava. Covid-19 and air pollution and meteorology-an intricate relationship: A review. *Chemosphere*, 263:128297, 2021.