The effect of COVID-19 and lockdown measures in Madrid's air quality by station.

Jorge Rodríguez Peña

Abstract—During 2020, the governments imposed mobility restrictions across the globe to stop COVID-19 from spreading. This project aims to evaluate the impact of these measures in Madrid's NO2 levels with the help of temporal and weather data. Using visual analytics, a linear regression model is built to predict values of NO2. Wind speed and pressure were the most relevant weather parameters, while hourly levels peaked during peak hours. The final model captured these variances and had a RMSE of 6.69 µg/m3. It was used to determine the impact of lockdown in NO2 levels across 24 stations in Madrid. The average reduction in NO2 levels was of the 53%. Stations at the north of Madrid, showed a higher impact than stations at the south of Madrid and the highest impact was observed in the station located at Madrid's airport. These results could be used by local authorities to improve the air quality in areas where mobility restrictions had higher impact, consequently improving the quality of life of the people that live there.

1 Problem Statement

The impact of lockdown during COVID-19 in 2020 was observed in pollution levels across the whole planet [1]. The decrease of China's pollution levels affected pollution levels across the globe [2]. Pollution is a significant factor in mortality in cities [3], a worrying fact that most of the time goes unnoticed. The lockdown pointed at one key factor of pollution levels: mobility.

The Spanish lockdown started on March 13th [4] and ended on June 21st, with a deescalation process that started on May 25th. On March 30th, the measures were reinforced and only essential workers were allowed to go to work (major lockdown).

This project focuses on the impact of lockdown in the city of Madrid's NO_2 levels, considering also the effect of weather. Weather variables like wind speed can affect the levels of pollution [7,8,9] and its effect on pollution levels should be considered to avoid overestimation of the impact of lockdown.

The EU considers that average NO_2 yearly emissions over 40 $\mu g/m^3$ are dangerous to the health of the citizens [5]. The information extracted from this project should help detect key areas of the city where mobility restrictions affected NO_2 levels in Madrid the most. Local authorities could further study the detected areas to seek for ideal solutions to improve air quality in the area.

2 State of the Art

The impact of mobility restrictions in pollution levels has been a subject of interest since the pandemic started, focusing on different pollutants including PM_{10}, SO_2, O_3 or, more frequently, NO_2.

Studies of pollution levels in Madrid during the national lockdown have been carried out. Baldasano (2020) [7] studied NO_2 levels in Madrid during March 2020, comparing them with the levels in March 2019 and March 2018 and using the wind speed registered by meteorological stations to explain differences. Briz-Redón et. al (2020) [8] modelled the levels of different pollutants across different cities in Spain (including Madrid) using a linear model to predict daily levels for each city. The first one uses linear plots to visualise the temporal evolution of NO2 levels. The second one uses heatmaps to visualise the variation in pollutant levels and boxplots to determine the marginal effect of the lockdown in each city. This use of visualisation focuses on explainability of results rather than as a tool to gain knowledge.

A neighbourhood study of NO2 levels in Madrid was carried out by Izquierdo et al. (2020) [10]. This study focused on the health impact of the Air Quality and Climate Change Plan for Madrid City approved in September 2017. It used choropleth maps to visualise the levels of different pollutants in 2012, a projected scenario for 2020, and the density of deaths associated to pollution to compare them to see if the implemented plan saved any lives.

A more visually engaging approach is the work carried out by Santos (2020) [9], where he studied the impact of lockdown in NO_2 emissions in London, using London's monitoring stations. The aim was to detect spatio-temporal patterns in 2019 and compare them to those observed in 2020. Patterns were sought seasonally, monthly, by day of the week and hourly with using heatmaps and scatterplots. He also used interpolation to plot heatmaps of NO_2 levels across London to better define areas with higher levels. This approach is distant from the locally approach taken in this project, that focuses only in the impact on each station.

This project's modelling approach is inspired by Mülbacher and Piringer (2020) [11]. The main idea is to use residuals as a new independent variable to search for patterns and areas where the model is failing. Then, the model can be split to model different areas or different sections depending on the distribution of the residuals. Heatmaps are also used to visualise the difference between models to determine where one model performs better than the other. All these approaches will be valuable to visualise the model results and improve it, to better capture the variation of NO_2 levels in Madrid.

As for the visuals, spatialization has been known as an excellent tool for constructing knowledge from data by transforming a high-dimensional dataset into lower-dimensional facilitating data exploration [13] and should be a relevant tool for the purposes of this project.

3 Properties of the Data

The data comes from the City of Madrid's Open Data Portal. This work will use the hourly air quality data for NO_2 [6] and weather variables [7] from March to June in 2019 and in 2020 collected by stations shown in Figure 1. shows the location of the stations across Madrid. The purpose is to study the impact of mobility restrictions comparing data from both years. The data was cleaned by imputing missing values as the average across all stations, and merged into a single dataset that contains:

- Date: Containing the date and hour the measurement was taken.
- Lockdown: Whether that date corresponds to the lockdown during 2020 or no.
- Station ID to identify where the data was collected.
- Geographical information of the station (i.e. latitude and longitude, shown in Figure 1).
- NO2 levels in µg/m3.
- Weather data: wind speed in m/s, temperature in ºC, relative humidity, barometric pressure in mb, solar radiation in W/m^2 and precipitation in L/m^2.

The data was cleaned to remove inconsistent data. To do so, boxplots of every variable were plotted. This process helped detect outliers in temperature (one value had a temperature bellow -40ºC), relative humidity (some values were negative, which means that a negative concentration of H_20_v was detected) and pressure (some values were bellow 900mb or 675mmHg). This process also showed the distribution of precipitation was skewed, with about 90% of the entries registering 0 L/m^2, suggesting that this variable will not be appropriate to the project.

Outliers were detected with Mahalanobis distance, and values 1.5 times greater than the interquantile range were removed. The final dataset contained about 96% of the original data.

The data spatialization used to colour the stations helps understand the differences observed in Figure 1., accomplishing its purpose of making sense of abstract information [12]. The data was spatialised using NO2 levels for both years, but more on that later. Stations coloured in a yellowish tone tend to be near big parks (suggesting the effect of vegetation in the pollution levels), while brownish stations located in the south and greenish stations in the north.

Figure 2. shows the correlation between each weather variable and the NO_2 levels. The figures suggest the most relevant predictors could be the wind speed and the barometric pressure, which have the highest absolute correlation among all stations with NO_2 levels. The station ID is coloured according to the spatialization colours used in Figure 1 and described in Figure 4.

4 Analysis

4.1 Approach

To determine the impact of the lockdown in each of the stations in Madrid, it is necessary to have a cleaned and well-structured data. Inconsistent data was removed using boxplots, and entry outliers were detected using the Mahalanobis distance. Outliers are defined as entries further than 1.5 times the interquantile range from the mean and will be removed. During this process, the correlation between weather variables and NO2 levels will be visualised with correlation heatmaps, both globally and by station, to select the most important variables to consider for model building. Temporal patterns and differences are also visualised to further understand the behaviour of NO_2 levels.

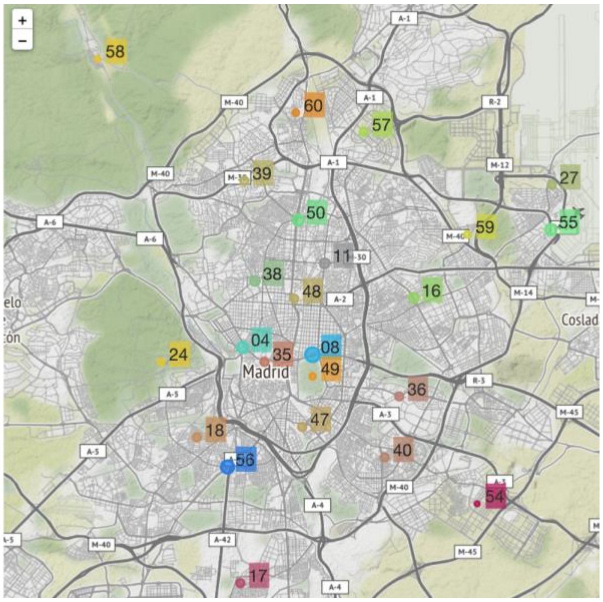

Figure 1. Location of the stations across Madrid. The size of each dot represents the observed variation from 2019 to 2020. The color of the station is calculated using spatialization.

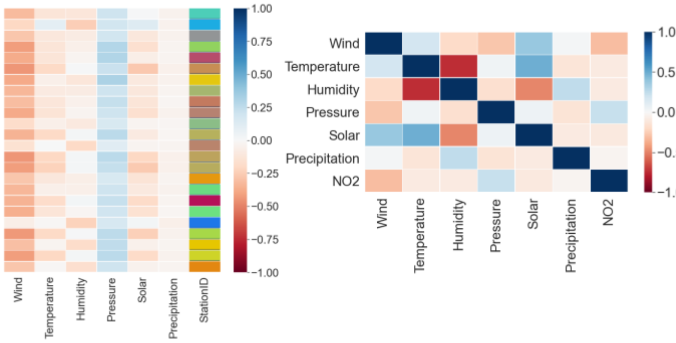

Figure 2. Left: Correlation between each weather variable and NO2 levels for each station. Right: Mean correlation across all stations between each weather variables and NO2 levels.

Since the purpose is to determine the effect in each station individually, spatialization provides a practical tool to represent each station individually while still addressing similarities and differences between stations. The spatialization will consider the mean values in each station during 2019 and 2020, as well as the differences between each year. To perform spatialization Principal Component Analysis (PCA) will be used. The results should help visualise similarities and differences between each station, adding knowledge to the study and adding that knowledge to the different plots that will be used.

To determine the effect of the lockdown, each station will be modelled using a linear model with ordinary least squares. The variable "Lockdown" will be used as an independent variable whose coefficient determines the effect of lockdown.

The modelling part will follow Mülbacher and Piringer's (2020) [11] approach, using residuals as a new independent variable and visualising their distribution across different variables. Using heatmaps and boxplots, this approach will help determine where is the model failing and how to solve that, using visual analytics to improve the computational approach.

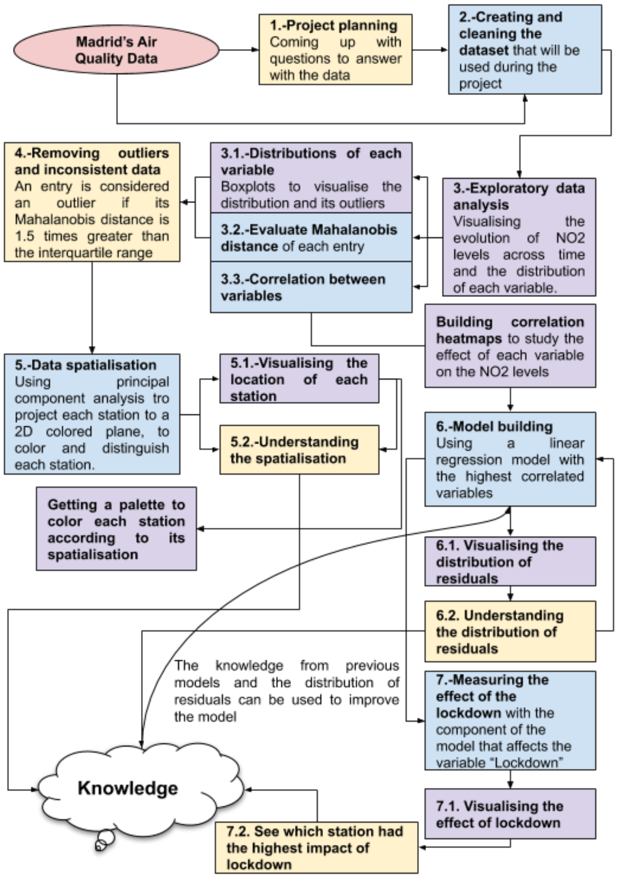
Madrid's Air Quality Data
1.-Project planning Coming up with questions to answer with the data
2.-Creating and cleaning the dataset that will be used during the project
4.-Removing outliers and inconsistent data An entry is considered an outlier if its Mahalanobis distance is 1.5 times greater than the interquartile range
3.1.-Distributions of each variable Boxplots to visualise the distribution and its outliers
3.2.-Evaluate Mahalanobis distance of each entry
3.3.-Correlation between variables
3.-Exploratory data analysis Visualising the evolution of NO2 levels across time and the distribution of each variable.
Building correlation heatmaps to study the effect of each variable on the NO2 levels
5.-Data spatialisation Using principal component analysis tro project each station to a 2D colored plane, to color and distinguish each station.
5.1.-Visualising the location of each station
5.2.-Understanding the spatialisation
6.-Model building Using a linear regression model with the highest correlated variables
Getting a palette to color each station according to its spatialisation
6.1. Visualising the distribution of residuals
6.2. Understanding the distribution of residuals
The knowledge from previous models and the distribution of residuals can be used to improve the model
7.-Measuring the effect of the lockdown with the component of the model that affects the variable "Lockdown"
Knowledge
7.1. Visualising the effect of lockdown
7.2. See which station had the highest impact of lockdown

Diagram. Diagram of the analysis approach. Yellow means human reasoning, blue means computational methods and purple means visualization.

The model will be fitted under two objectives: it must be as simple as possible and as accurate as possible. Once the model is fitted, the effect of the lockdown for each station can be measured and compared. At the end of the project, stations with a higher impact of the lockdown will be determined. The results should provide useful information to improve the air quality of that area. The diagram above provides a summary of the analysis approach.

4.2 Process

The initial results show that only about 3.8% of the entries were outliers of the Mahalanobis distance. Figure 3 shows outliers are evenly distributed across the distribution of NO_2 and that they represent a minimal effect on the data and therefore can be removed.

The cleaned data was used to spatialise the stations using the mean value of each station during 2019 and 2020, as well as the difference between each year and the fraction of the mean value in 2019 that the difference represents, i.e.:

$$\frac{2019-2020}{2019}$$

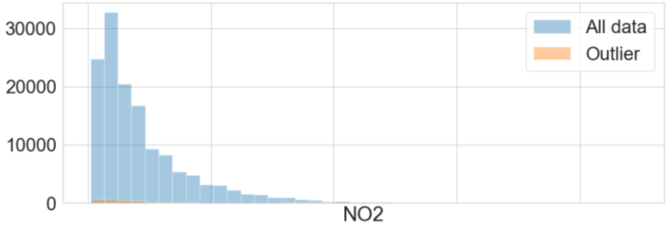

Figure 3. Histogram of the distribution of outliers compared with the distribution of the data.

The results of the spatialization can be seen in Figure 4. The axis in Figure 4. depicts the tendency of the data. The location of the stations is shown in Figure 1 and the spatialization marks a spatial pattern across stations in Madrid. It appears that stations 56 and 8 have the highest NO_2 levels and the highest difference after the lockdown, with station 56 having a lower fraction than station 8. These stations are located near to the city centre and in major crossroads.

Near to the origin of the scatter plot, in brown, stations from the south of Madrid are located. Bellow that cluster, in green, stations from the north of Madrid are plotted. This separation of south and north probably is related to the way of living of their inhabitants. Neighbourhoods in the south of Madrid are highly populated and are the home of workers that have to commute or drive to their working sites, while the north of Madrid has richer, less-populated neighbourhoods.

Finally, on the bottom left of the scatterplot, stations in yellow represent stations near to parks or green areas that are affected by the vegetation that surrounds them.

To address the temporal variation of NO_2 levels, two plots were created. The first one is shown in Figure 5. It contains hourly values of NO_2 during 2019, 2020 and the observed difference between those years. In the background, lockdown is plotted as a rectangle, specifying where the major lockdown and the deescalation took place. A clearly drop in the NO_2 levels can be observed in 2020 once the lockdown starts, while in 2019 some stations are above 200 $\mu g/m^3$. In particular, the bluish stations from the spatialization have higher values during 2019, while yellowish stations have lower values. This pattern is observed in the other two plots, but the values of bluish stations are much lower in the second one.

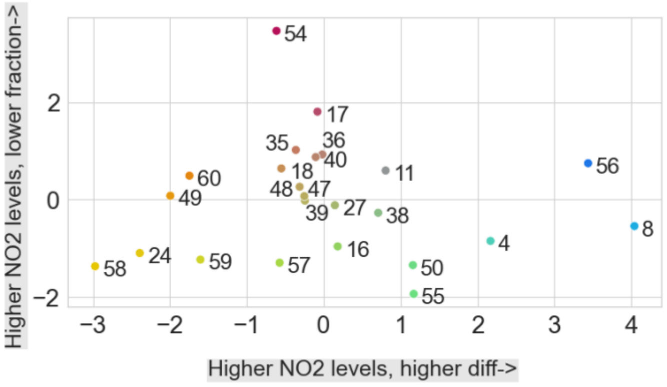

Figure 4. Spatialization of the stations using PCA. The axis are labelled with the tendency of the data across each component.

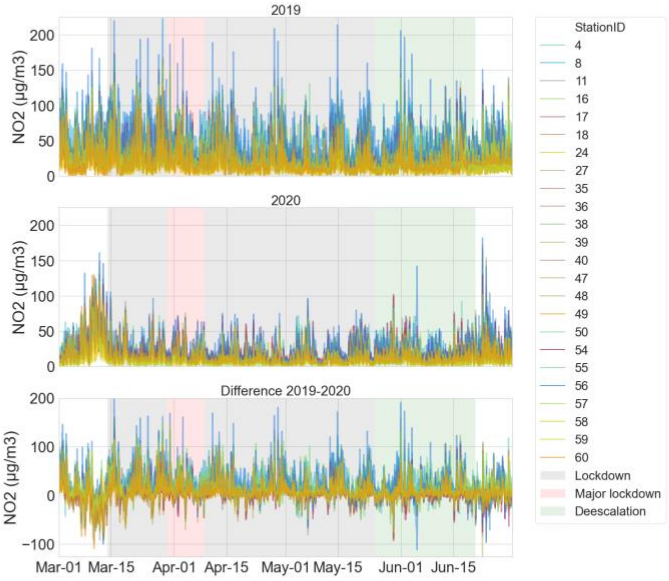

Figure 5. Hourly evolution of NO_2 levels across different stations. The first two plots have values from 2019 and 2020 respectively, while the third plot contains the observed difference during the same day between the two years. Behind, the presence of lockdown (with major lockdown and deescalation specified) is added as a squared background.

The second plot contains mean hourly values for each day of the week for each station with and without the lockdown and is shown in Figure 6. The difference between the two periods is plotted too. The first two heatmaps show a severe time dependence with peak hours: values increase during the first part of the day (going to work) and during the evening (coming from work). This pattern is also observed in Figure 9 (bottom right). Values at the weekend are also much lower than values during weekdays. The third plot shows that the higher differences were observed during these periods of peak hours, suggesting the effect of the mobility restrictions. These heatmaps suggest NO_2 levels are highly correlated with hourly values and therefore should be considered while modelling.

As for weather variables, the correlation with NO_2 levels was already shown in Figure 2., with pressure and wind being the highest correlated variables. The negative correlation for wind speed implies that higher values of wind speed mean lower values of NO_2, which makes sense considering that wind can carry the pollutant's particles away. Higher pressure means higher levels of NO_2, which can be explained by acknowledging that higher pressures push NO_2 particles in higher atmospheric levels to lower ones. The rest of the weather variables are significant to certain stations, except for precipitation which has a very low correlation across all stations.

The first model built for each station used all weather variables except for precipitation and Lockdown as independent variables and had a root-mean-square error (RMSE) of 10.14 µg/m^3.

Visualisation of residuals showed the model tended to under-predict by far higher values of NO_2. This was due to the skewness of the distribution of NO_2. As shown in Figure 7. a "more normal" distribution was obtained by taking the natural logarithm of the NO_2 levels.

Instead of modelling the NO_2 levels, the natural logarithm of the NO_2 levels was modelled obtaining this time a RMSE of 9.57 µg/m^3. Visualisation of residuals in Figure 8. (top) showed the model was failing to capture temporal evolution. Under-predicting the month of march and over-predicting Saturdays and Sundays. Peak hour values had also a notable RMSE.

To solve this problem, the variable month was transformed to a dummy variable. The RMSE dropped to 9.27 µg/m3.

After that, the difference between weekends and weekdays was still very high so a model was built for each day of the week, dropping the RMSE to 8.63 µg/m3. But residuals during peak hour and night hours were still significantly high.

To fix this, a study of the hourly residuals was carried out. Figure 9 (top) shows the pattern observed in residuals. The pattern seemed sinusoidal and the residuals were modelled using a Fourier transform to obtain their principal frequencies.

The Fourier transform was carried out across all stations. After visualising the results for all stations, six frequencies appeared five times or more as the top five frequencies in all of them. These are shown in Figure 9 (bottom left) and in the table below, with an approximation of the frequency and the periodicity represented by that frequency:

Value	Approximation	Periodicity
0.265	$\pi/12$	24h
0.463	$5\pi/34$	13.6h
0.529	$\pi/6$	12h
0.794	$\pi/4$	8h
1.058	$\pi/3$	6h
1.323	$5\pi/12$	4.8h

Figure 9 (bottom right) shows the fitted residuals after using the six frequencies, amplitudes and phases obtained in the Fourier transformation.

Using cosines with the obtained frequencies and phases as new independent variables, a fourth model was fitted. The RMSE dropped to 6.69 µg/m^3. Figure 8 (bottom) shows the residuals' distribution. It is an improvement from the second model (Figure 8 (top)), although some regions still show problems. In particular, the region of low wind speed values and hours between 11 and 17. After and inspection of the data, wind speeds of those values for the specified hours were unusual, resulting in the errors observed in that zone.

A test set was used to prevent overfitting. It had a RMSE of 6.97 µg/m^3, proving the validity of the model.

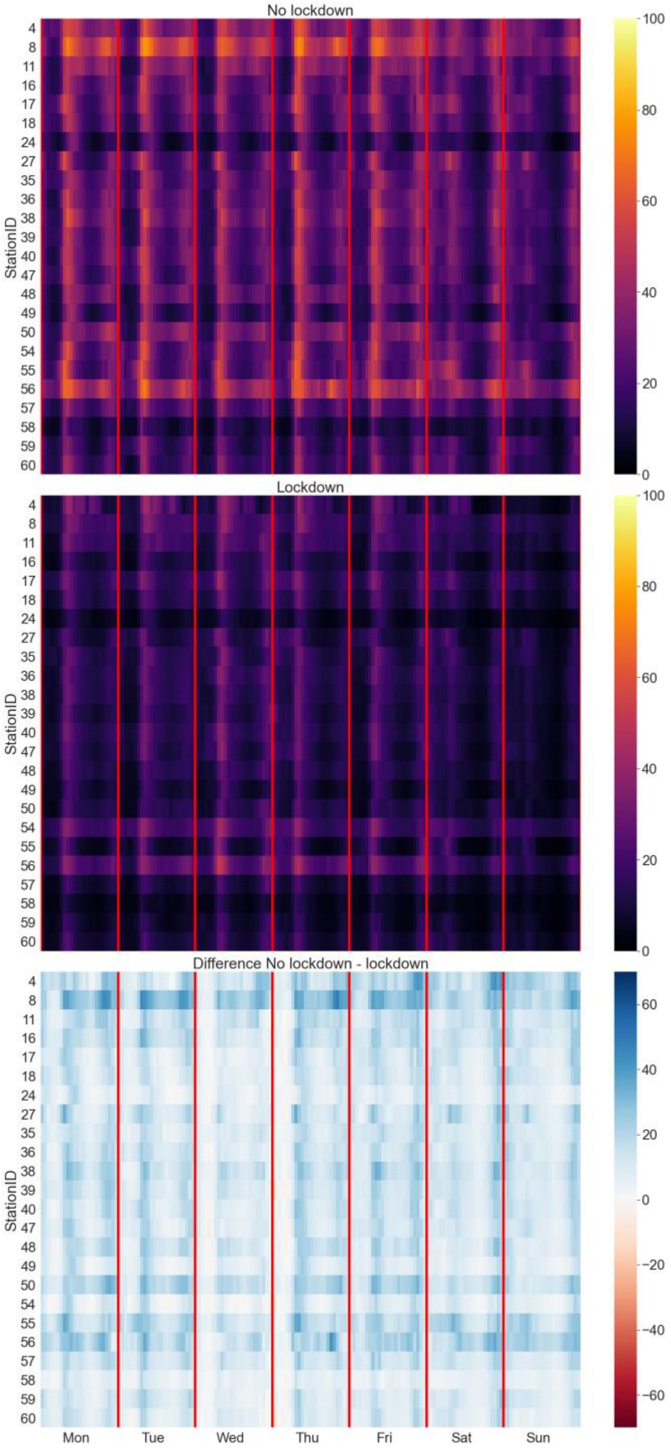

Figure 6. NO_2 levels by hour by day of the week. The first two plots contain values from periods without and with lockdown (respectively), while the last plot contains the observed difference.

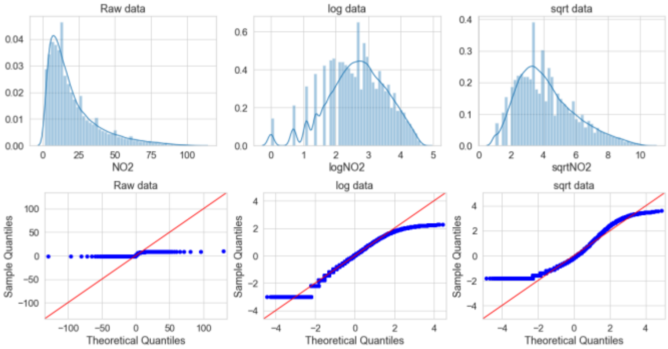

Figure 7. Different distributions of NO_2 levels. From left to right: raw values, logged values, square-rooted values. Bellow, the q-q plots to compared the distributions against a normal distribution.

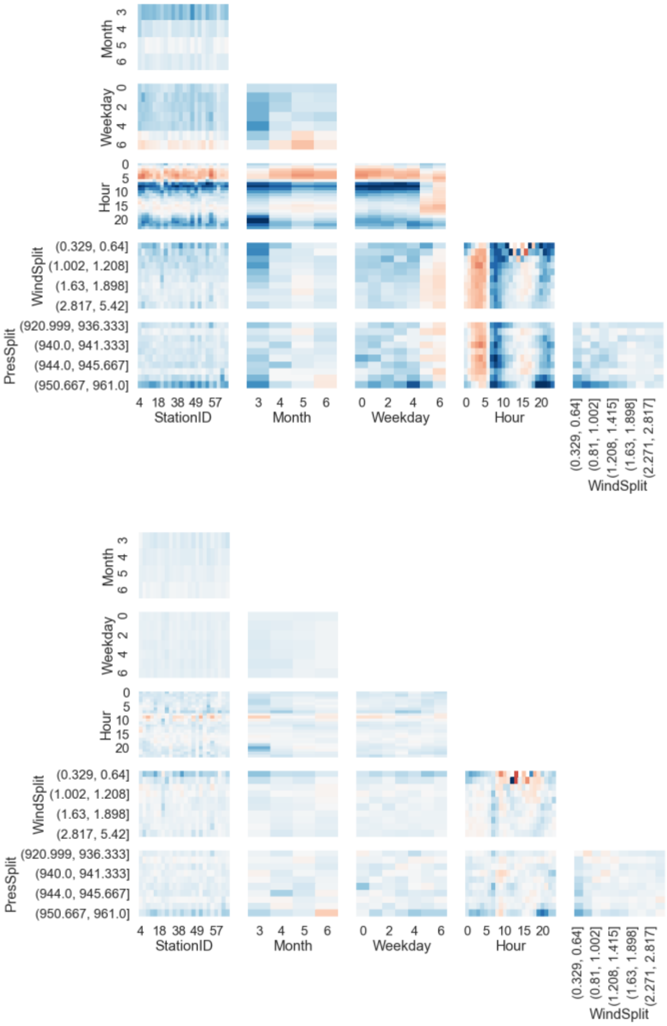

Figure 8. Top: Distribution of residuals obtained for the second model. Bottom: Distribution of residuals obtained for the final model.

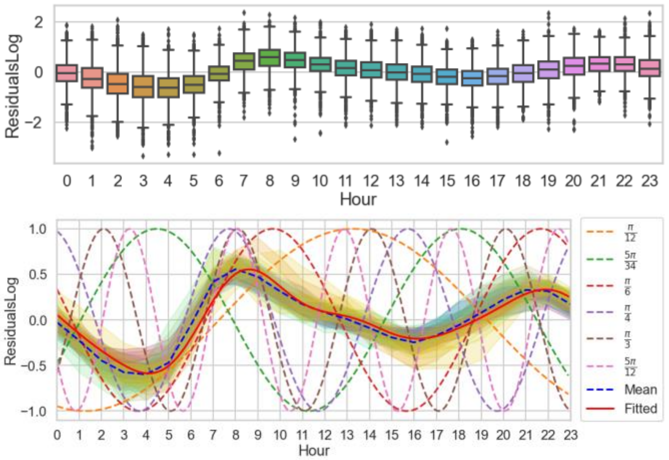

Figure 9. Top: Hourly distribution of residuals for the third model. Bottom left: Number of times each frequency appeared in the top five frequencies of each station. Bottom right: Fitted mean residuals values with the 6 top frequencies compared with the hourly residuals for each station, also plotted in this plot.

With these results, the model is simple enough for explainability and has enough accuracy to generate valuable predictions. Further modelling was carried out but the improvement of the model was insufficient compared to the increase in complexity, so this was considered the final and optimal model.

Once the model is fitted, the coefficient of the effect of lockdown can be extracted. The model uses as a dependent variable the logarithm of the NO$_2$ levels, wich means:

$$\ln y_{NO_2} = C_{lockdown} + f(time, weather)$$

Where $C_{lockdown}$ is the coefficient of the variable “Lockdown” and $f(time, pressure, wind\ speed)$ the linear combination of the other variables. Therefore, to obtain the NO$_2$ levels it is necessary to exponentiate the function, getting:

$$y_{NO_2} = e^{C_{lockdown}} \cdot e^{f(time, weather)}$$

So, the exponential of the coefficient of the variable “Lockdown” modulates the intensity of the NO$_2$ levels obtained by fitting the model with wind speed, pressure and the Fourier components for hourly values. If $e^{C_{lockdown}}$ is greater than one, that means that lockdown increases the NO$_2$ levels, otherwise, it means that lockdown decreases the levels.

4.3 Results

The effect in each station can be visualised in Figure 10. For each station the 95% confidence interval for the effect of lockdown was bellow 1. NO$_2$ levels decreased a 53% on average.

Station 54 had the lowest impact. This station is located further from the city centre (but near to the A-3, an important highway) in one of the poorest neighbourhoods in Madrid and is not a main destiny (including the city centre) rather an origin of mobility. Station 55 had the highest impact. This station is located at Madrid’s Barajas Adolfo Suárez Airport. Without any flights, pollution levels dropped significantly in this station, also adjacent to the countryside. The impact of lockdown was higher in northern Madrid (greenish areas) than in southern Madrid (brownish areas). The model does not explain these differences but they could be explained with the mobility of the workers to their respective workplaces, considering the economic differences between the south (poorer) and the north (richer).

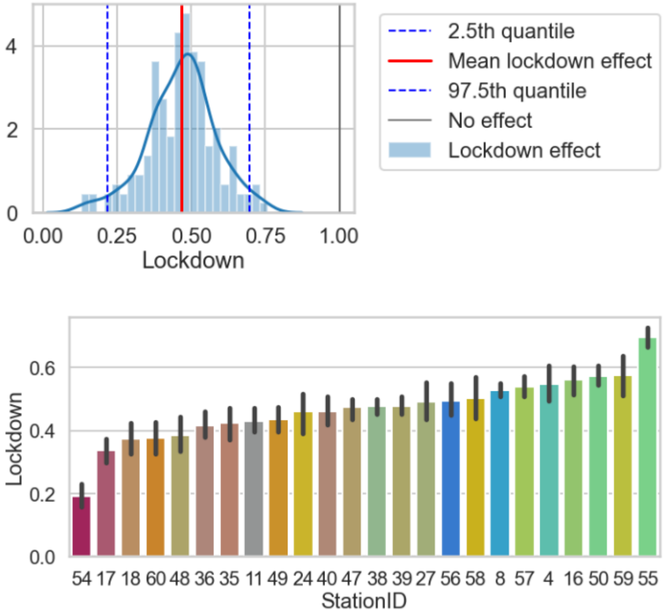

Figure 10. Top. The mean effect of lockdown for all stations. **Bottom.** Stations ordered by impact of lockdown. The height of the bar represents the total reduction of NO2 levels.

5 Critical reflection

Visual analytics was used as a tool for model-building and pattern recognition. The implementation of spatialization helped visualise differences between the north and the south (greenish and brownish stations), as well as detecting stations located in the middle of big parks (yellowish stations). Keeping this spatial distribution in mind, it is easier to interpret the results and visualise differences.

Building a consistent and simple model was made clear by looking at the distribution of residuals across variables. That helped visualise areas where the model was failing, and improved it by modelling each area separately (a model per station per month and per day of the week) or by generating new temporal variables that captured the variability of the target variable (with the computational help of a Fourier transformation).

The weather variables were, in general, unhelpful with only Wind and Pressure having a high correlation with NO$_2$ levels for all stations. The model heavily relied on temporal variables: a different model was generated for each month and day of the week and the frequencies obtained from the Fourier transformation for hourly values. Prove of that is that the highest decrease in the RMSE was observed after modeling the hours. This suggests that other variables can be used to better model the pollutant levels. One such variable can be traffic congestion.

The final model helped visualise and determine which stations were more affected by the lockdown than the others. With the help of the palette obtained with spatialization, the results showed that northern stations were more affected than southern stations. However, the model failed to capture why.

One possible explanation is that, as said before, the model is not considering other variables like the traffic density. Spatio-temporal patterns in traffic density can be used to compare them with the observed patterns in NO_2 levels. This could confirm the theory that the density of traffic is higher in the city centre and in the north due to the mobility of people in the south towards their workplaces. A study of the direction of wind in the city of Madrid could also be used to check if the wind comes from the north and goes to the south usually. This could be another possible explanation on why the southern stations were not as affected as northern stations.

The model helped determine the effect of lockdown in each station and this information could be used by local authorities to improve the air quality in the areas around the stations. Improving the air quality is an improvement in the quality of life and a reduction of mortality related to pollution.

References

The list below provides examples of formatting references.

[1] Muhammad, S., Long, X., Salman, M., 2020. COVID-19 pandemic and environmental pollution: a blessing in disguise? Sci. Total Environ., 728 https://doi.org/10.1016/j.scitotenv.2020.138820.

[2] Wang, Q., Su, M., 2020. A preliminary assessment of the impact of COVID-19 on environment - a case study of China. Sci. Total Environ. 728, 38915. https://doi.org/10.1016/j.scitotenv.2020.138915

[3] Renjie Chen, Haidong Kan, Bingheng Chen, Wei Huang, Zhipeng Bai, Guixiang Song, Guowei Pan, on Behalf of the CAPES Collaborative Group, Association of Particulate Air Pollution With Daily Mortality: The China Air Pollution and Health Effects Study, *American Journal of Epidemiology*, Volume 175, Issue 11, 1 June 2012, Pages 1173–1181, https://doi.org/10.1093/aje/kwr425

[4] Real Decreto 463/2020, de 14 de marzo, por el que se declara el estado de alarma para la gestión de la situación de crisis sanitaria ocasionada por el COVID-19. Boletín Oficial del Estado, March 14th 2020, num 67, p. 25390-25400.

[5] EU, 2008, Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe (OJ L 152, 11.6.2008, p. 1–44)

[6] Madrid's Air Quality Service. *Calidad del Aire. Datos horarios años 2001 a 2020.* 2020. https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=f3c0f7d512273410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default

[7] Madrid's Air Quality Service. *Datos meteorológicos. Datos horarios desde 2019.* 2020. https://datos.madrid.es/sites/v/index.jsp?vgnextoid=fa8357cec5efa610VgnVCM1000001d4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD

[8] Baldasano, José. (2020). *COVID-19 lockdown effects on air quality by NO2 in the cities of Barcelona and Madrid (Spain).* Science of The Total Environment. 741. 140353. 10.1016/j.scitotenv.2020.140353.

[9] Briz-Redón, Álvaro & Belenguer-Sapiña, Carolina & Serrano-Aroca, Ángel. (2020). Changes in air pollution during COVID-19 lockdown in Spain: A multi-city study. Journal of Environmental Sciences. 101. 10.1016/j.jes.2020.07.029.

[10] Santos, Laurent Jose & Haworth, James. (2020). *Impacts of covid-19 on urban air pollution in London.* 10.13140/RG.2.2.15144.80643.

[11] Rebeca Izquierdo, Saul García Dos Santos, Rafael Borge, David de la Paz, Denis Sarigiannis, Alberto Gotti, Elena Boldo, *Health impact assessment by the implementation of Madrid City air-quality plan in 2020, Environmental Research,* Volume 183, 2020, 109021, ISSN 0013-9351, https://doi.org/10.1016/j.envres.2019.109021

[12] Mühlbacher, Thomas & Piringer, Harald. (2013). A Partition-Based Framework for Building and Validating Regression Models. IEEE transactions on visualization and computer graphics. 19. 1962-71. 10.1109/TVCG.2013.125.

[13] Skupin, A. and I. Fabrikan, S., n.d. *Visualisation: A Geographic Approach*. 1st ed. pp.61-79.

[14] Mühlbacher, Thomas & Piringer, Harald. (2013). A Partition-Based Framework for Building and Validating Regression Models. IEEE transactions on visualization and computer graphics. 19. 1962-71. 10.1109/TVCG.2013.125.

See the full Project at:
https://github.com/jorgerodpen/MadridPollution