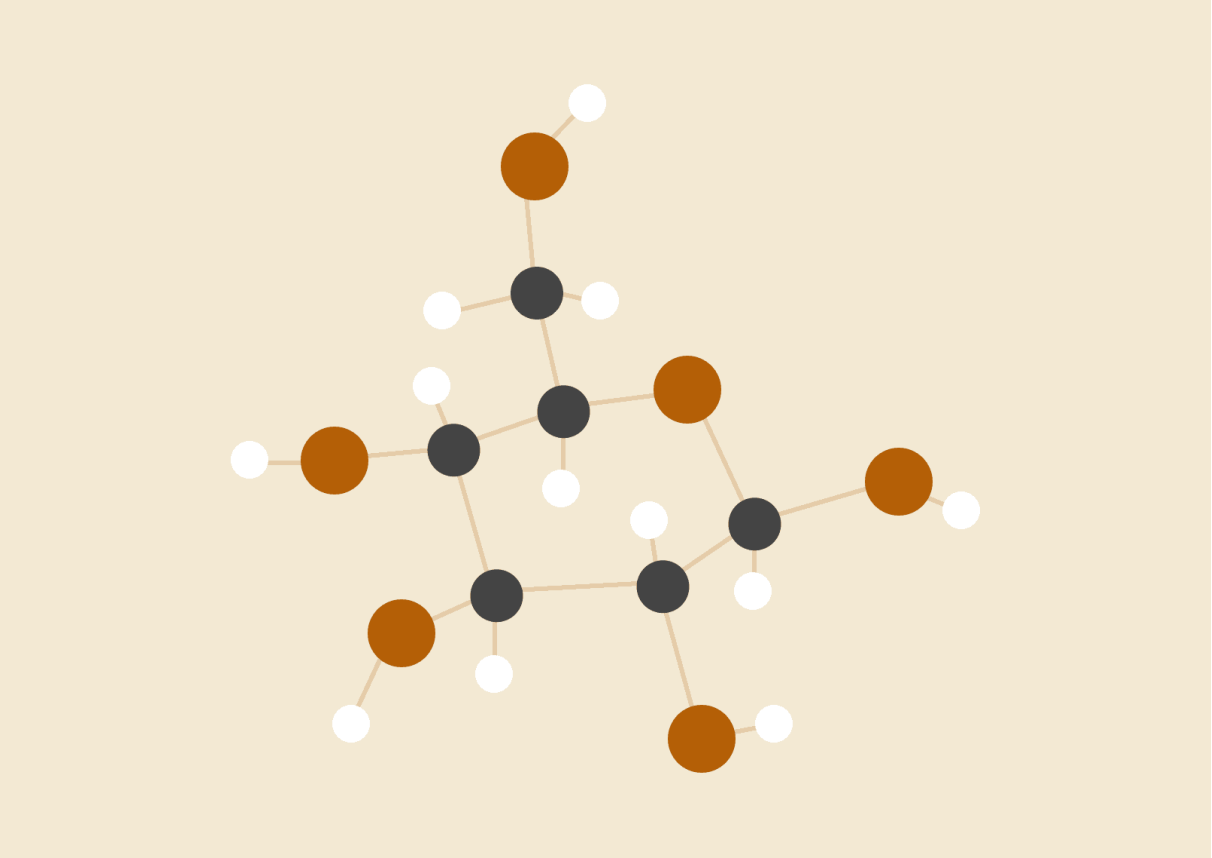
Madrid's Air Quality

Explanatory Analysis and Forecasting

ANH-THU, BUU, EYA, GWENAELLE



I- Introduction

Global warming is the gradual increase in temperature of the Earth's atmosphere, primarily due to greenhouse effects caused by the level of gas like carbon dioxide, azote, and other pollutants. Therefore, it is one of the main consequences of increasing contamination. For decades, the environmental association and scientists have tried to alert the general public and the government to this problem. However, contamination is raising many other issues that imperil the sustainability of human beings. For example, the **greenhouse effect** retains the heat between the atmosphere and the Earth. So, it is disturbing the ecosystem, and the biosphere components, so it impacts agriculture. It also implies that humans are directly impacted because our food comes from agriculture. Besides, since gas emissions cause heat, respiratory disease risks also increase.

In 1972, the states decided to concretize the principle of **sustainable development** and decision-making by creating the sustainable development goal and the **conferences of the Parties** (COP). During these conferences, the governments discuss the solutions to reduce the increasing global temperature. Our dataset deals with air pollutant measures in Madrid between 2001 and 2018. The other dataset is about the station descriptions. The 3808224 values were measured by different stations in the city from Kaggle. All dataset is composed of 19 variables:

- **Date**: from 2001 to 2018 (2018 are incomplete)
- **Stations**: stations id
- **BEN**: benzene level measured in μg/m³. Benzene is an eye and skin irritant, and long exposures may result in several types of cancer, leukaemia and anaemias. Benzene is considered a group 1 carcinogenic to humans by the

IARC.

- **CO**: carbon monoxide level measured in mg/m³. Carbon monoxide poisoning involves headaches, dizziness and confusion in short exposures and can result in loss of consciousness, arrhythmias, seizures or even death in the long term.
- **EBE**: ethylbenzene level measured in μg/m³. Long term exposure can cause hearing or kidney problems and the IARC has concluded that long-term exposure can produce cancer.
- **MXY**: *m*-xylene level measured in μg/m³. Xylenes can affect not only air but also water and soil, and a long exposure to high levels of xylenes can result in diseases affecting the liver, kidney and nervous system (especially memory and affected stimulus reaction).
- **NMHC**: non-methane hydrocarbons (volatile organic compounds) level measured in mg/m³. Long exposure to some of these substances can result in damage to the liver, kidney, and central nervous system. Some of them are suspected to cause cancer in humans.
- **NO2**: nitrogen dioxide level measured in μg/m³. Long-term exposure is a cause of chronic lung diseases, and are harmful for the vegetation.
- **NOx**: nitrous oxides level measured in μg/m³. Affect the human respiratory system worsening asthma or other diseases, and are responsible for the yellowish-brown color of photochemical smog.
- **OXY**: *o*-xylene level measured in μg/m³. See MXY for xylene exposure effects on health.
- **O3**: ozone level measured in μg/m³. High levels can produce asthma, bronchitis or other chronic pulmonary diseases in sensitive groups or

outdoor workers.

- **PM10**: particles smaller than 10 μm. Even though they cannot penetrate the alveolus, they can still penetrate through the lungs and affect other organs. Long term exposure can result in lung cancer and cardiovascular complications.
- **PXY**: *p*-xylene level measured in μg/m³. See MXY for xylene exposure effects on health.
- **SO2**: sulphur dioxide level measured in μg/m³. High levels of sulphur dioxide can produce irritation in the skin and membranes, and worsen asthma or heart diseases in sensitive groups.
- **TCH**: total hydrocarbons level measured in mg/m³. This group of substances can be responsible for different blood, immune system, liver, spleen, kidneys or lung diseases.
- **PM2.5**: particles smaller than 2.5 μm level measured in μg/m³. The size of these particles allow them to penetrate into the gas exchange regions of the lungs (alveolus) and even enter the arteries. Long-term exposure is proven to be related to low birth weight and high blood pressure in newborn babies.
- **NO**: nitric oxide level measured in μg/m³. This is a highly corrosive gas generated among others by motor vehicles and fuel burning processes.
- **CH4**: methane level measured in mg/m³. This gas is an asphyxiant, which displaces the oxygen animals need to breathe. Displaced oxygen can result in dizziness, weakness, nausea and loss of coordination.

To get a precise idea of why Madrid was startled by the European Union and how the council breaks the dead blocks, we first convert the measurements to the Common Air Quality Index (CAQI), an air quality index used in Europe since 2006.

Then we will use the monthly average of this index to see the behavior of forecasting the CAQI if the government would not act. Then we will test the efficiency of the environmental laws promulgated by Spain's capital by forecasting the index after the law applications.

II- Methodology

1. Dealing with the missing data

While going through the data, we notice a lack of information in the old stations. However, it is not the case that the city council is not well aware of this matter. On the contrary, They probably knew but refused to fix it for some particular reason.

A short data exploration reveals some exciting details about the stations. We can check for missing data during the whole period and see them side by side to get a better perspective on the activity of each station. Using the library **missingno**, we get a fast and easy-to-understand visualization of when the data is present in each station.

2. Data pre-processing

2.1. Reformatting the structure of the dataset

The dataset contains stations at the highest hierarchical level: each station history can be individually extracted from the file for further study. Inside each station's dataset, all the particle measurements that such station has registered from 01-2021 to 05-2018. Not every station has the same equipment. Therefore each station can measure only a specific subset of particles. There are 19 separate pieces of data as input: 18 CSV data for each year (2001 - 2018) and one XLS data for stations. We bind them together since we need a general overview of the air pollutants in Madrid. We mainly use **forecast**, **tidyverse**, **ggplot2**, and **xts** libraries from R to perform our analysis.

2.2. Choice of the variables

In order to highlight the multi-level impact of the contamination in Madrid, we choose three variables that are mainly responsible for human health, such as cancers. They are **Particles smaller than ten μm** (PM10) and **Ground-level Ozone** (O3). Besides, we selected **Nitrogen dioxide** (NO2), which consistently impacts the environment.

The European Union directives assess hourly values of NO2, daily average values of PM10, and 8- hour average values for O3 in addition to a range of year average criteria. The hourly calculation is done for reasons of attractiveness outlined above. Since 2006, the CAQI index has been calculated for hourly, daily, and yearly averaged data. The calculation is mainly based on three pollutants of significant concern: PM10, NO2, and O3.

3. Forecasting

3.1. ARIMA model

An Auto Regressive Integrated Moving Average (ARIMA) is a statistical analysis model that uses time-series data to better understand the dataset or predict future trends. It is a method for forecasting or predicting future outcomes based on a historical time series. It is based on the statistical concept of serial correlation, where past data points influence future data points.

An ARIMA model can be understood by outlining each of its components as follows:

- Autoregression (AR): refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.
- Integrated (I): represents the differencing of raw observations to allow for the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).
- Moving average (MA): incorporates the dependency between an observation

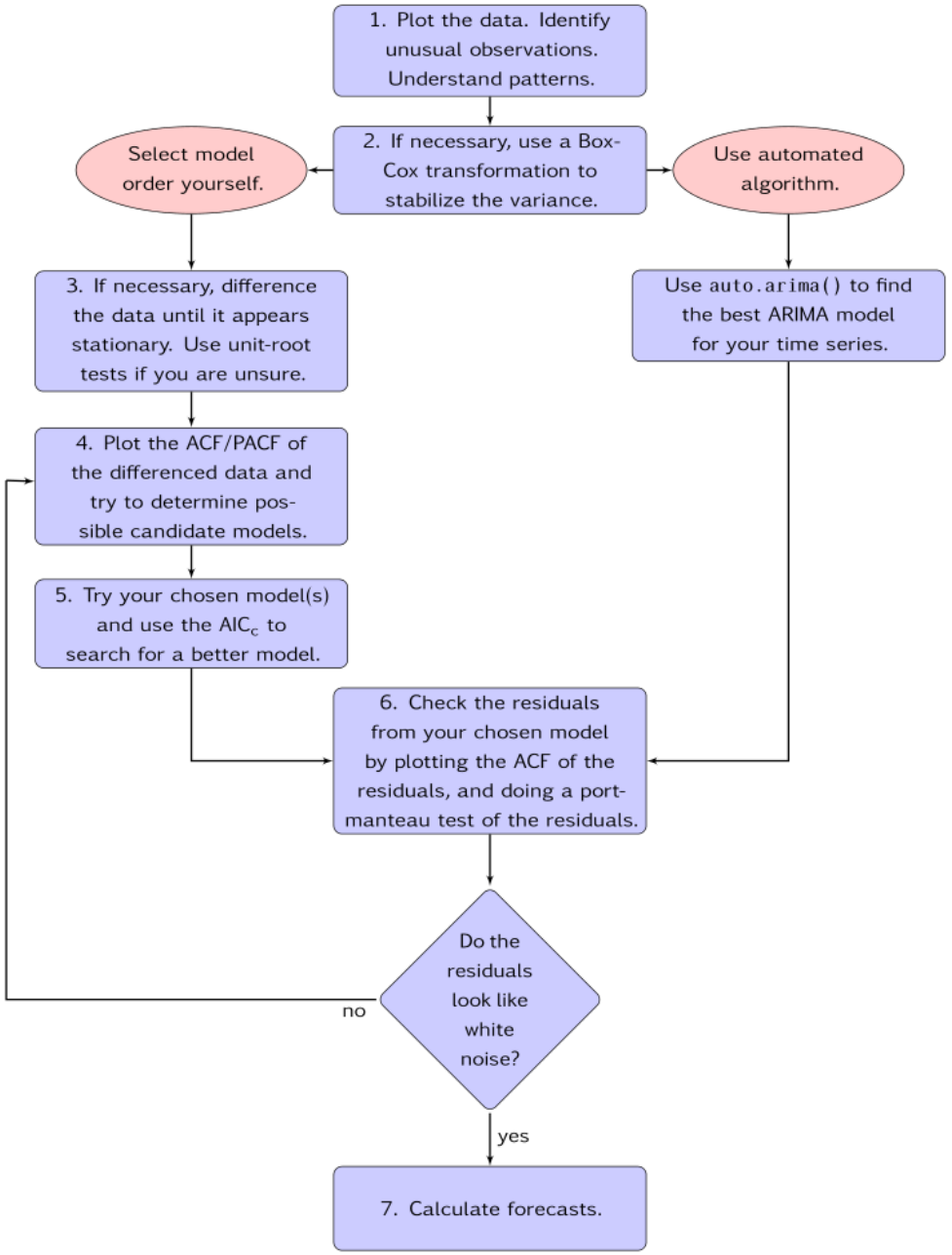
and a residual error from a moving average model applied to lagged observations.

An Seasonal ARIMA model is labeled as an ARIMA(p, d, q)(P, D, Q)$_m$, wherein:

- **p** is the number of autoregressive terms.
- **d** is the number of differences.
- **q** is the number of moving averages.
- **P** is the **seasonal** number of autoregressive terms.
- **D** is the **seasonal** number of differences.
- **Q** is the **seasonal** number of moving averages.
- **m** is the number of observations per year.

3.2. Auto Arima in R

We use the function **auto.arima** from the **forecast** library to generate our prediction model. The output returns an ARIMA model with the best set of hyperparameters to fit our data. Then we use the **autoplot** function to visualize the forecasting.


1. Plot the data. Identify unusual observations. Understand patterns.
2. If necessary, use a Box-Cox transformation to stabilize the variance.
Select model order yourself.
Use automated algorithm.
3. If necessary, difference the data until it appears stationary. Use unit-root tests if you are unsure.
Use auto.arima() to find the best ARIMA model for your time series.
4. Plot the ACF/PACF of the differenced data and try to determine possible candidate models.
5. Try your chosen model(s) and use the AIC$_c$ to search for a better model.
6. Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals.
Do the residuals look like white noise?
no
yes
7. Calculate forecasts.

III- Results

1. Missing data

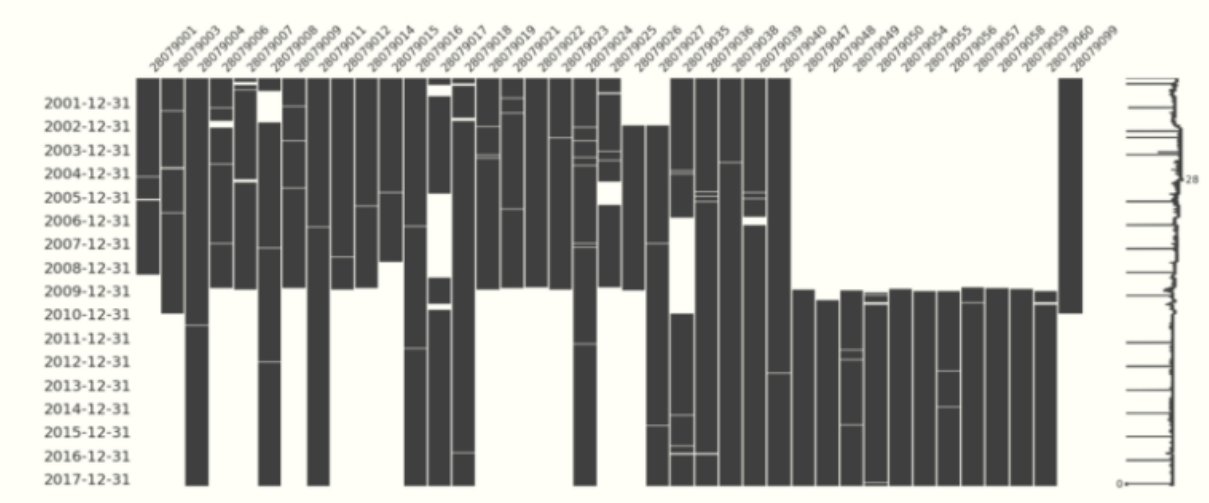


Figure 1: Missing data in each station during the period 2001-2018

In Figure 1, missing values are represented by gaps. The columns are the stations' ID and the rows are the date of record. Thus, the missing values here are regarded as the inactivity of the station for a given day. We can denote that almost half of the stations ceased their activity. By looking at the stations data set, we can see that only 6 stations span all the period. Therefore, it is the same station with a different ID.

2. Explanatory analysis

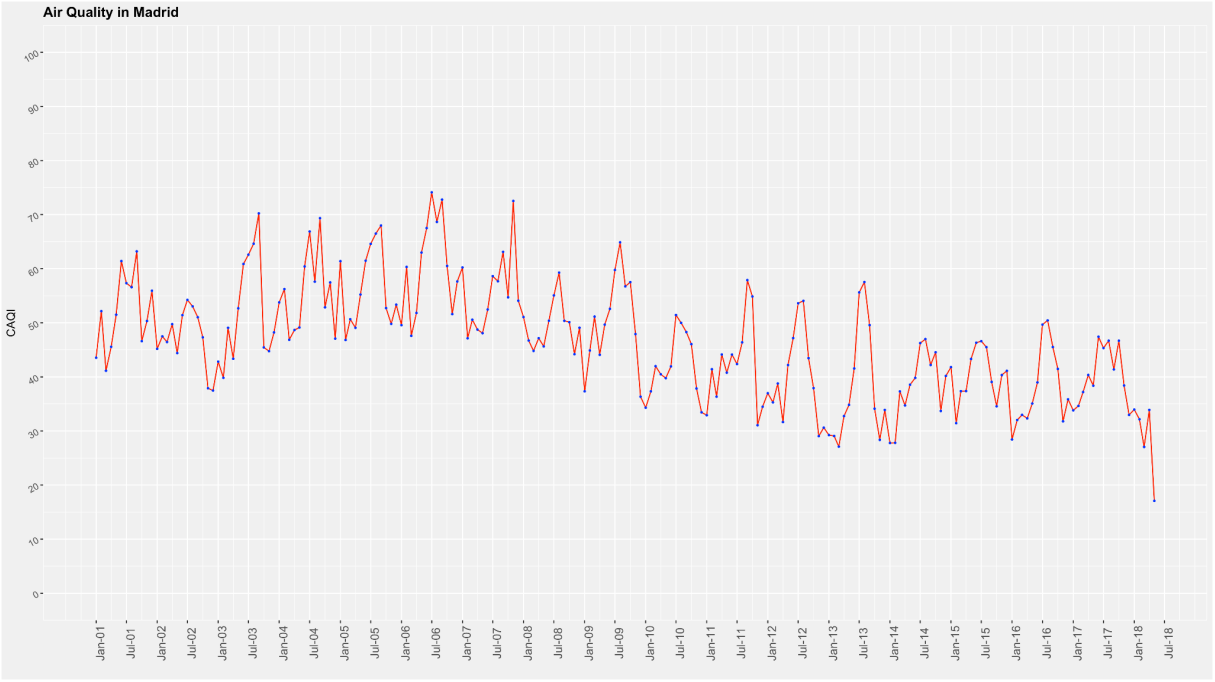


Figure 2: Monthly Moving Average of Madrid's Air Quality

In Figure 2, there are two trends over time. First, at the beginning of the timeline, we see that pollutants tend to increase. Then, in July 2006, the values of the CAQI decreased. Therefore, we can say that overall, the index values in Madrid are decreasing over time.

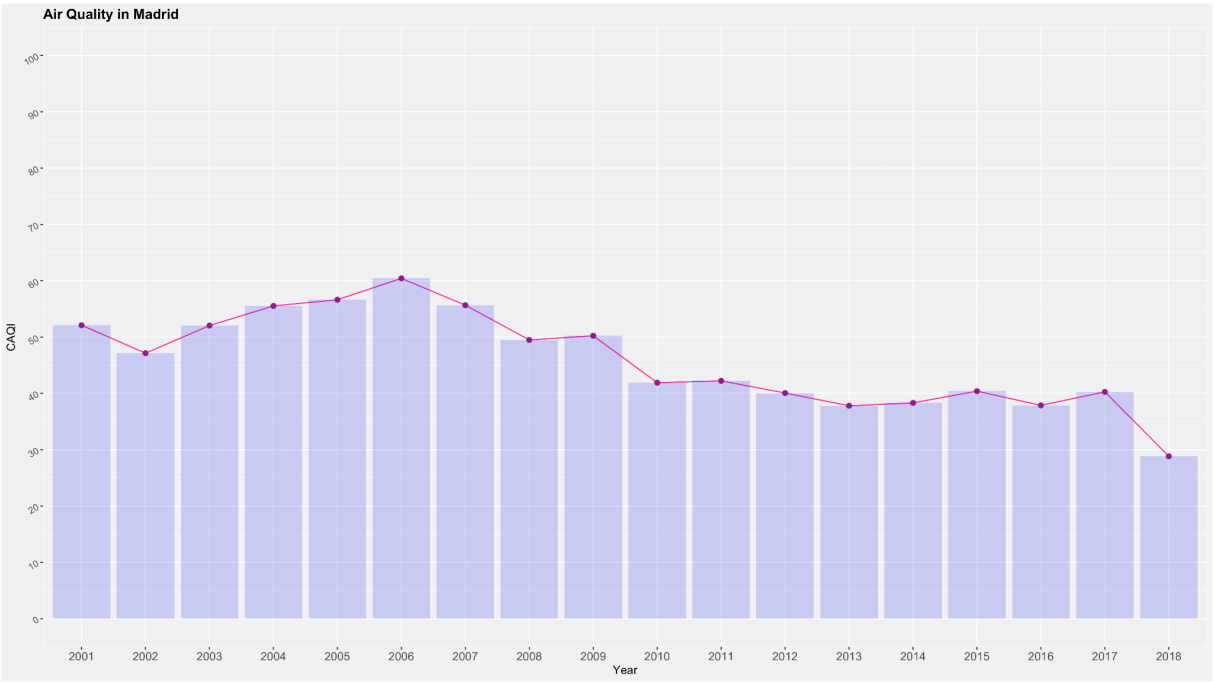


Figure 3: Yearly Moving Average of Madrid's Air Quality

In Figure 3, we can see the yearly average of the CAQI index. This bar plot helps us to describe more precisely the variation of trend we have just mentioned. We can clearly see that the CAQI values are increasing from 52 to 60 on average, between 2001 and 2006. Then, there had been a gradual decrease during 12 years. We can identify two stages on the slight drop from the graph. The period from 2006 to 2008 (from 60 to 49), and from 2009 to 2017 (from 50 to 40).

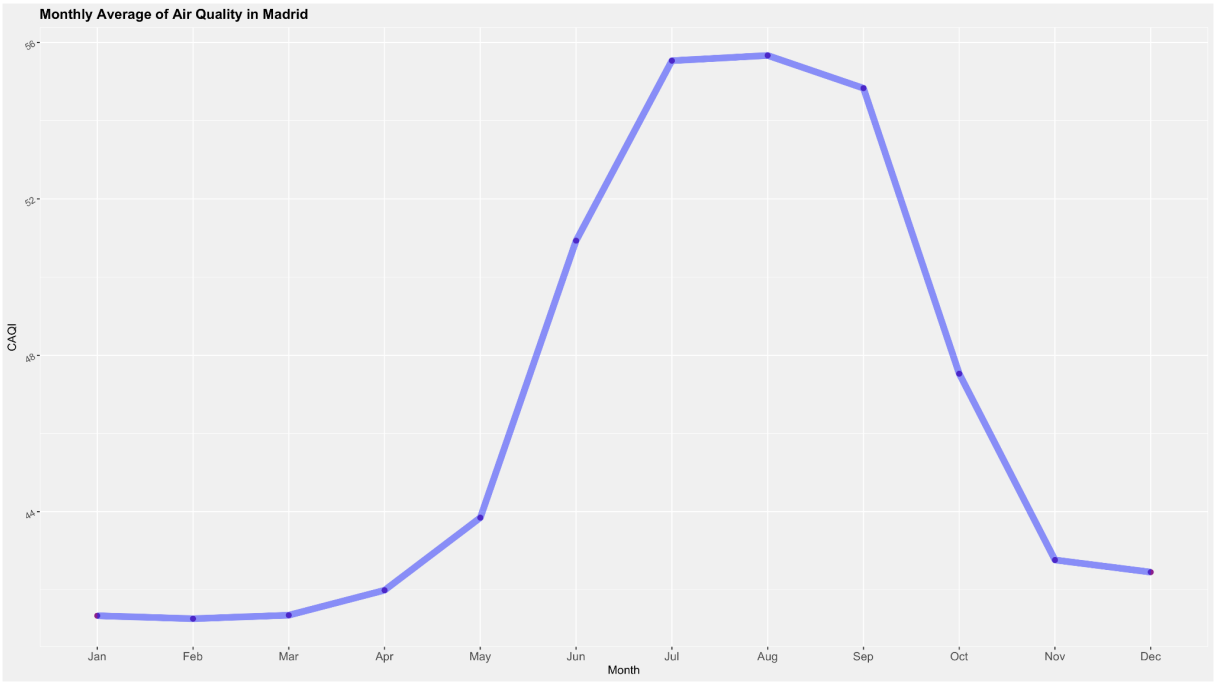


Figure 4: Monthly Moving Average of Madrid's Air Quality

In Figure 4, we can see the information about Madrid's monthly average air quality from 2001 to 2018. From that, we can quickly know in which month the pollution increases or decreases. It is very clear from the overall trend that the pollution increases in summer from the beginning of May and decreases in September and October. According to the graph above, the monthly average made a significant climb around 7 CAQI from May (44 CAQI) to June (51 CAQI). In August, the contamination hit the highest point for nearly 56 CAQI. In September, the total pollution suddenly goes down from 55 CAQI to 47 CAQI in October and decreases in the following months.

3. Forecasting

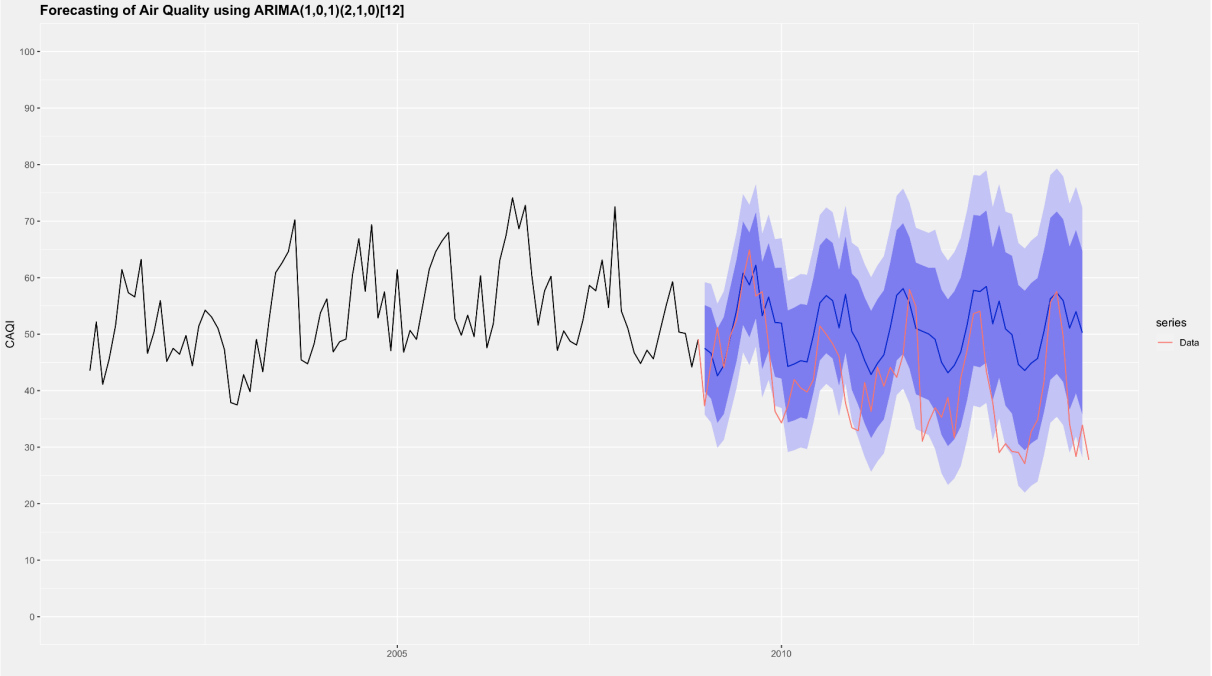


Figure 5: Forecast of Madrid's Air Quality from 2009

Figure 5 presents the results of Madrid's Air Quality forecasts obtained by applying our model ARIMA $(1, 0, 1)(2, 1, 0)_{12}$ for five years from January 2009 to January 2014. The red line represents the actual data from 2009, and the blue line represents our forecast. Both lines have a decreasing trend. However, the forecast line decreases slower. Furthermore, the peak in each year of the actual data is lower than the forecast of about 5 AQI.

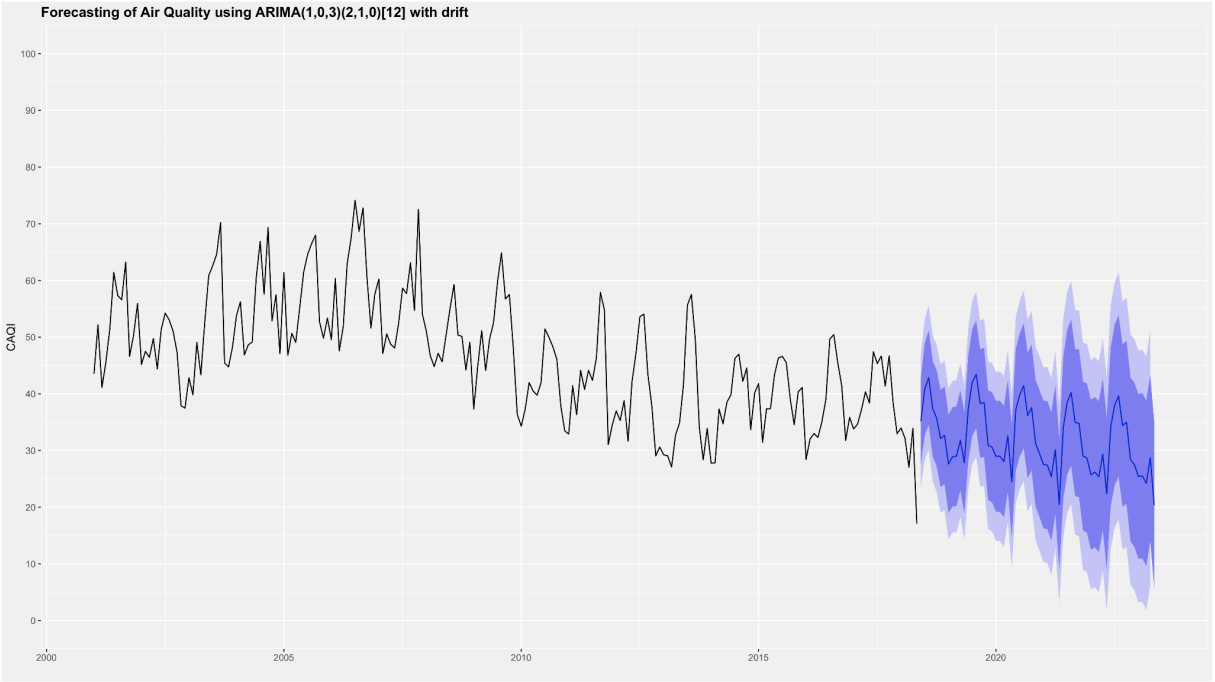


Figure 6: Forecast of Madrid's Air Quality in next 5 years (2023)

Figure 6 presents the results of Madrid's Air Quality forecasts obtained by applying our model ARIMA $(1, 0, 3)(2, 1, 0)_{12}$ for five years from June 2018 to June 2023. From this forecast, we see that in the next five years, the Air Quality of Madrid will have a decreasing trend with a mean of around 30 CAQI, and its peak will be around 41 CAQI.

IV- Discussion

1. Explanatory Analysis

Even if the city reduced the number of pollutants in 2009, Madrid still broke all contamination records and endured fierce criticism for its immobility. No regulation of the most polluting vehicles in the city center and a lack of green spaces across it were casting doubt on the city's sustainability. That same year, the European Union recommendations went from a friendly recommendation to compulsory, and most of the limits were not being fulfilled. This can explain the change in the general trend we denoted in Figure 2 and Figure 3. By looking at the Council environmental plan of Madrid, we can see that the city invested in ads to promote electric cars and promulgated laws to reduce the CO emission in the city by limiting the number of cars and the speed limitation.

Moreover, we notice the seasonality in Madrid's Air Quality data. As a matter of fact, there is a rise and fall in the CAQI that regularly repeats over the same period (As seen in Figures 2 and 3). We know that Madrid, like many other cities in Spain, is attractive for tourism during the high season. Hence, the increase in CAQI from June to September in Figure 4 may be partly due to the population rising during the summer vacation.

2. Forecasting the Air Quality of Madrid

Within the EU, 2009 was a year during which efforts to mitigate climate change were consolidated. Based on the forecast result, we can see that the actual data is much lower than the forecasting we got by applying the ARIMA model. In short, the climate action of the EU had an impact on the data of Air Quality in Madrid.

Based on the available data and the forecast, the air quality in Madrid has improved during the last decade due to government policy measures on air pollution. The concentrations of primary air pollution decreased drastically after 2009 and continue to have a decreasing trend for the next five years. These observations clearly indicate that the developments in EU environment policy during 2009 and the Madrid environment plan have been efficient in improving air quality. Nevertheless, in terms of data reliability this topic should be discussed. Indeed, before 2009, the measurement was done in certain places. We have figured out that the Madrid government has changed the station location to green spots or simply closed some of them which makes the data not accurate anymore. By knowing that we should remember the fact that to perform a good analysis forecasting comparison between two different situations we need to have the same or common input as basic. If we don't do so, it could be assumed that the accuracy of the analysis would suddenly drop, thus from the following statement we think that the result is less optimistic. We can not deny the fact that a lot of effort has been done by the government in order to improve the environment situation, but we also have to admit that the situation may not be as good as on paper.

V- Conclusion

To sum up, we saw that the air pollution in Madrid was terrible until 2009. Before 2009, almost all the pollutant concentrations had been higher than European's recommendations. Besides, we saw that the predictions without a reaction from the state were warmful. After hardening the environmental laws in Europe, we highlighted a significant decrease in the air quality index, which can be explained by the city's investment in the environment. Thus, we could conclude that according to the forecasting, the results of the plan were encouraging.

Nevertheless, after studying some articles about the subject, we found that the city council's reaction was ambiguous. Instead of taking action against the pollution, their efforts focused on masking the results. In late 2009, most stations were moved to different locations or closed. One of the closed stations registered 74 µg/m3 of NO2 when the limit recommended by the European Union was 42µg/m3. 18 out of 24 stations registered levels higher than this limit. The stations were moved to more propitious locations: stations located in crossings and roundabouts were moved to greener areas, which registered lower measurements when the levels are peaking. Even most surviving stations changed which substances were being measured, which can be seen in the data.

References

CAQI Index: http://www.airqualitynow.eu

ARIMA: https://otexts.com/fpp2/arima-r.html

Environmental plan:

Madrid suspende en contaminación | Madrid

Madrid environmental plan

Limitation of the data:

Madrid no consigue reducir la contaminación | Madrid | EL PAÍS

http://www.elmundo.es/elmundo/2011/02/04/ciencia/1296820221.html

Environment Policy Review:

https://ec.europa.eu/environment/archives/pdf/policy/EPR_2009.pdf