Time series analysis and forecasting of $N0_2$ and PM_{10} daily mean concentration levels at Eixample, Barcelona[1]

Adrián Hinojosa-Calleja[2]

Abstract

We analyze the NO_2 and PM_{10} concentration registers collected by an atmospheric pollution measure station at the Eixample neighborhood in Barcelona, Spain. We evaluate the level of short exposure to those pollutants between June 2019 and July 2022 according to the Air Quality Guidelines established by the World Health Organization. By means of an Auto Regressive Integrated Moving Average method we construct a one step forecast time series model to predict the NO_2 and PM_{10} daily mean concentration levels.

Keywords: air quality, time series analysis, ARIMA model, prediction of pollutants' levels

1. Introduction

The intense industrial activity, the high volume of traffic, the Mediterranean climate with low average rainfall over the course of the year and the arrival of Saharian dust clouds favors the concentrations of polluting elements in the breathing air of Barcelona. This has a negative impact for the human health.

The two most harmful elements and more abundant air pollutants in Barcelona are Nitrogen Dioxide (NO_2), an irritant gas that attacks the lungs and is mainly been produced by internal combustion engines, and the Particulate Matter of less than 10 microns (PM_{10}), which especially affects the respiratory and cardiovascular systems and that are produced in the processes of combustion, driving and breaking of vehicles, construction works, or dust that arrives from the Sahara [1].

[1]This work is presented as a final project for getting the Data Science, IT Academy's certificate. `https://www.barcelonactiva.cat/es/itacademy`

[2]IT Academy, Carrer de Roc Boronat, 117, 127, 08018 Barcelona, Spain. Email: hinojosa.a@gmail.com.

The World Health Organization (WHO) has recently upgraded its Air Quality Guidelines (AQG) for long and short term exposure of different pollutants [8]. We analyze the hourly data collected by a measuring station at the Eixample neighborhood in Barcelona with two main objectives:

- To evaluate the level of short exposure to NO_2 and PM_{10} pollutants according to the WHO's AQG between June 2019 and July 2022.
- To study a one step forecast ARIMA model which predicts NO_2 and PM_{10} daily mean concentrations based in the values from previous days.

2. State of Art

There exists in the literature several works which study time series forecasting of air contaminants in different places around the world. We do not plan to be exhaustive on its presentation, we briefly mention some selected examples which make use of an ARIMA model.

The main source of inspiration for the present work is [6] which presents an ARIMA model for CO, NO_2, O_3 and $PM_{2.5}$ concentrations in Sofia, Bulgaria between 2015 and 2019. The trends in $PM_{2.5}$ concentrations of Fuszhou, China between August 2014 and July 2016 were studied using ARIMA in [9]. Numerical air quality forecasts in Hong Kong were improved using ARIMA models in [5].

3. Materials and Methods

The data is collected by one of the stations of the Catalonia's Vigilance and Forecasting of Atmospheric Pollution Net, located at the Eixample neighborhood in Barcelona [2]. The data ranges from 1 July 2019 to 31 June 2022, providing values for NO_2 and PM_{10} pollutants in micrograms per cubic meter ($\mu g/m^3$).

3.1. Data Preparation

The initial raw data contains hourly measurements (see Figure 1). Due to the presence of missing values, imputation of data is performed.

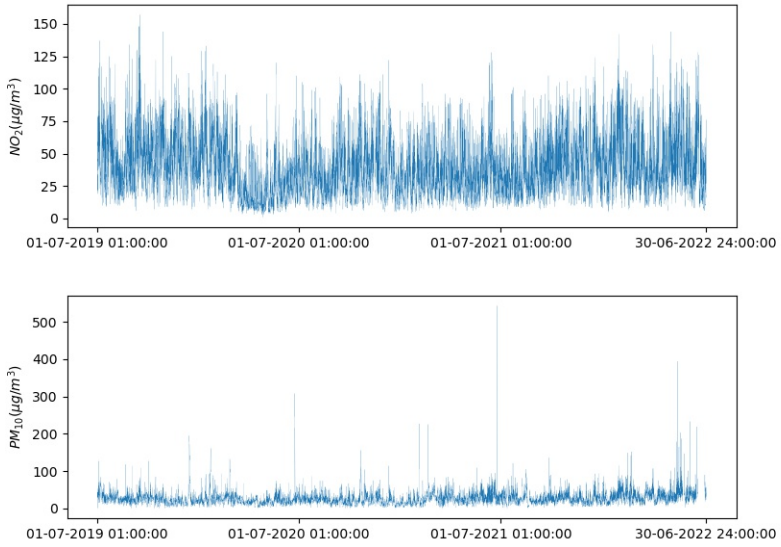

Figure 1: NO_2 and PM_{10} concentrations measured every hour from 1 July 2019 to 30 June 2022.

Polluter	**Mean**	**Std**	**Min**	**Max**	**Missing**
N02	39.76	21.67	1	157	3.67%
PM10	26.59	16.89	1	543	6.10%

Table 1: Quality of the raw data set by pollutant

3.1.1. Quality of Raw Data

The raw data sets are explored to check their quality. The results are summarized in Table 1. Missing values are present due to technical incidences during the operation of the station.

3.1.2. Imputation of Missing Values

Since for the chosen method in time series analysis, it is essential to avoid missing vales, they are handled in the raw set. Similarly to [3], the missing values are completed by using the mean of the corresponding time series.

4. Analytic Methods

ARIMA model is applied to perform the data analysis. It consists of three components: Autoregressive (AR), Integrated (I) and Moving Average (MA) models (see [7] for example).

4.1. ARIMA Model

The main parameters of the method are:

- Number of previous observations p.
- Degree of diferencing d.
- Size of moving average q.

The AR (p) model obtains p previous observations as follows:

$$y_t = \mu + \sum_{i=1}^{p} \varphi_i y_{t-i} + e_t$$

where y_t is the prediction for some t and y_{t-i} defines p previous observations of the same time series, φ_i denotes the regression coefficients, μ is a constant and e_t is a normal random variable.

The MA (q) performs forecasting relying on moving averages of the past random error terms as follows:

$$y_t = \mu + \sum_{i=1}^{q} \theta_i e_{t-i}$$

where θ_t represents the regression coefficients, q is the order of moving average and $(e_{t-i})_{i=1,\dots,q}$ is a white noise.

The ARIMA (p, d, q) model can be defined as follows:

$$w_t = \mu + \sum_{i=1}^{p} \varphi_i w_{t-i} - \sum_{i=1}^{q} \theta_i e_{t-i}$$

where $w_t = \Delta^d y_t$ and Δ is the backward difference operator. For example,

$$\Delta^2 y_t = \Delta(y_t - y_{t-1}) = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}).$$

4.2. Evaluation Metrics

We make use of the following metrics to evaluate the forecast:

- Mean Absolute Error (MAE) defined by:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|x_t - y_t|.$$

- Mean Absolute Percentage Error (MAPE) which expresses the accuracy of a model as a percentage of error:

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\frac{|x_t - y_t|}{|x_t|}.$$

In both cases, n is the number of samples used for fitting the model, x_t the actual value and y_t the predicted value a time t.

4.3. Software Libraries

For model stabilization and identification, the Hyndman-Khandakar algorithm for automatic ARIMA modeling is used [4], which is available in the *forecast* package in R as the *auto.arima* function, to select the best parameters (p, d, q) based on the Akaike Information Criterion (AIC): the lowest AIC models are selected as the best models. Once the best parameters are identified, we make use of the *arima* function for perform a one step forecasting and compute the evaluation metrics.

5. Experimental Results

5.1. Air Quality Levels

The WHO has established recommendations of AQG levels for classical pollutants. The exceed of the AQG levels is associated with important risks to public health. Additionally, Interim targets aim to promote a shift from high air pollutant concentration, with acute and serious health consequences, to lower concentrations [8].

Table 2 contains the Interim targets and the AQG levels for 24h mean concentrations of NO_2 and PM_{10}.

Pollutant	Interim target				AQG level
	1	2	3	4	
$NO_2\left(\mu g/(m^3 \times 24h)\right)$	120	50	-	-	25
$PM_{10}\left(\mu g/(m^3 \times 24h)\right)$	150	100	75	50	45

Table 2: Air Quality Guidelines for 24h mean exposure to NO_2 and PM_{10} particles.

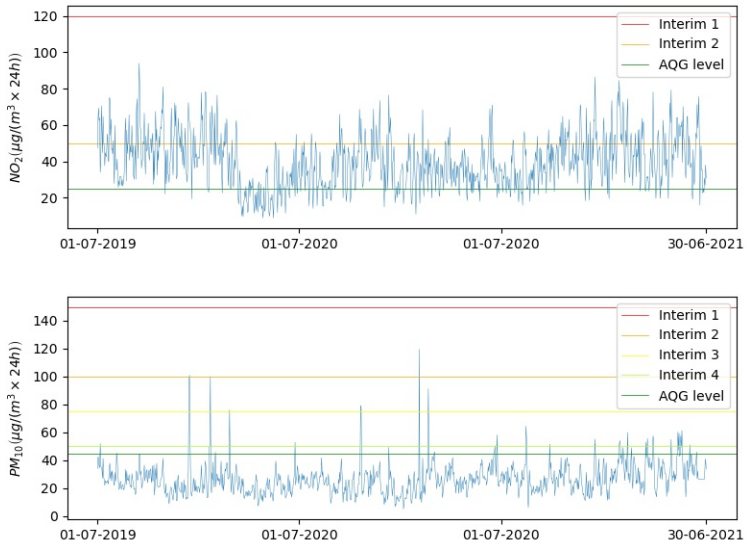

Figure 2: 24h mean of NO_2 and PM_{10} concentrations from 1 July 2019 to 30 June 2022

After imputing the missing values we use the data set with hourly registers for computing the 24h concentration mean for every pollutant (See Figure 2). In both cases there is not any register that exceeds the upper limit value of the Interim 1 stage.

In Figure 3 we present the percentage of days per trimester between July 2019 and June 2022 when the NO_2 concentration levels have been on each of the categories established by the WHO. NO_2 exhibits an unusual low level of concentration between March 2020 and November 2021 attributable to the decrease of activities during the COVID 19 pandemic. After that period we observe a recovery of the previous values and slight trend of air quality improvement.

In Figure 4 we present the percentage of days between July 2019 and June 2022 per trimester

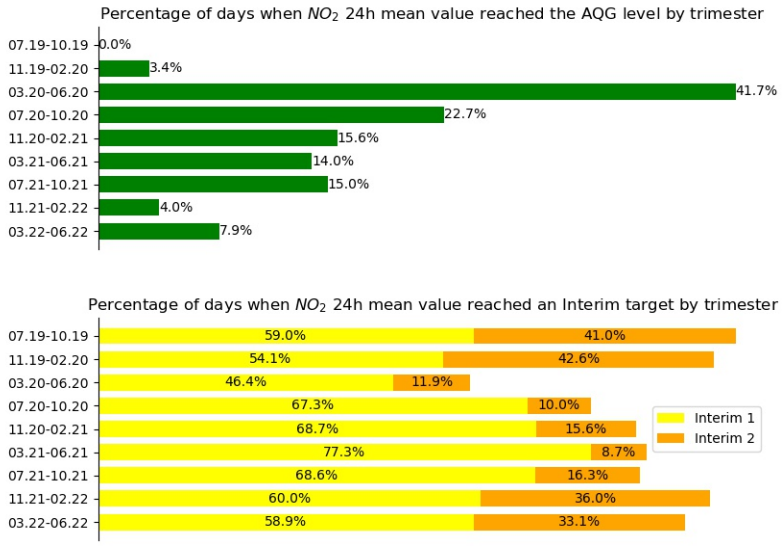

Figure 3: Percentage of NO_2 24 h mean concentration levels on each of the WHO's categories by trimester from 1 July 2019 to 30 June 2022

when the PM_{10} concentration levels have been in each of the categories established by the WHO. On the analyzed period PM_{10} levels exhibits a slight trend of diminishing air quality.

5.2. ARIMA Models

Before fitting the parameters of the ARIMA model, we divide the NO_2 and PM_{10} daily mean time series on train and test sets. We make use of a 80%-20% splitting ratio thus the train runs from 1 July 2019 to 26 October 2021, the test from 27 October 2021 to 30 June 2022.

The *auto.arima* function is executed for fitting the parameters and coefficients of the model by means of the train set. Then one step forecasting is performed over the train and test sets. Table 3 contains the selected fitted parameters by *auto.arima* and the values of the evaluation metrics between the train, test and the one step forecast obtained by the model (see Figure 5).

In the PM_{10} forecast the MAPE metric evaluates a better performance on the test set than in the train. This is due to a gap of missing values in the raw data between 02 May 2022 and 28 June 2022. Since the mean value were imputed in that period, the time series takes a constant value

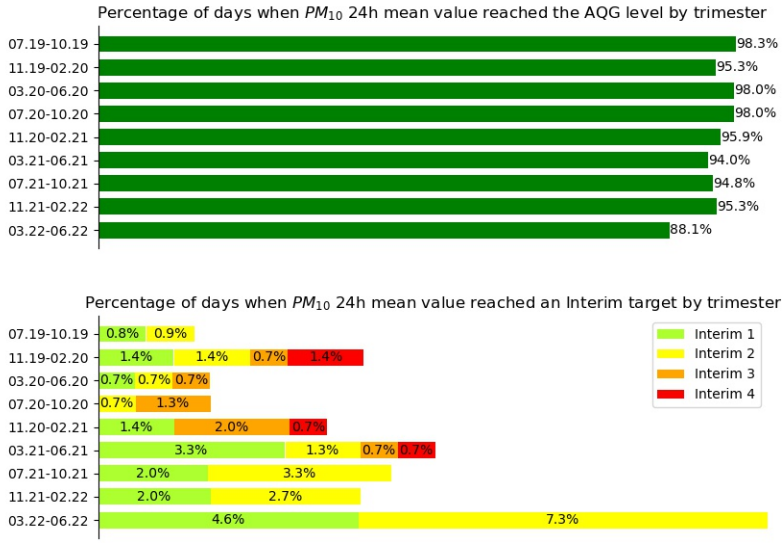

Figure 4: Percentage of PM_{10} 24 h mean concentration levels on each of the WHO's categories by trimester from 1 July 2019 to 30 June 2022

Pollutant	**p**	**d**	**q**	**MAE train**	**MAPE train**	**MAE test**	**MAPE test**
NO_2	2	1	1	8.32	24.96%	10.48	26.72%
PM_{10}	1	0	0	5.54	24.77%	6.86	22.69%

Table 3: MAE and MAPE of NO_2 and PM_{10} for the choosen parameters p,d,q on the train and test set one step predictions.

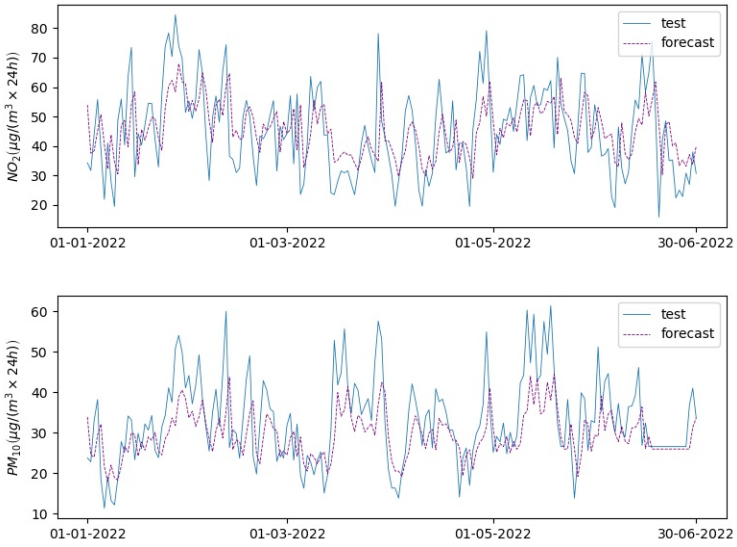

Figure 5: ARIMA model with one step predictions for NO_2 and PM_{10} 24h means on the test set from 01 January 2022 to 30 June 2022.

which the model predicts with high accuracy.

6. Conclusions

Even they exhibit very different behaviors NO_2 and PM_{10} daily concentrations still have not reached the 100% AQG level goal. The corresponding authorities should propose and ensure the compliance of realistic low and long future objectives in terms of the Interim targets for progressively increasing Barcelona's air quality.

We propose a one step forecast model which according to the MAPE metric in both pollutants has an average error of less than the 30%. This can be useful to protect the population for exposure in advance in case of high pollution episodes.

References

[1] Air quality. What is happening to the air we breathe. Avaiable online: `https://ajuntament.barcelona.cat/qualitataire/en/qualitat-de-laire/what-happening-air-we-breathe`

(access on 18 July 2022).

[2] Air quality data from the measure stations of the city of Barcelona. Avaible online: `https://opendata-ajuntament.barcelona.cat/data/en/dataset/qualitat-aire-detall-bcn` (access on 11 July 2022).

[3] Goyal, P., Chan, A. T., & Jaiswal, N. (2006). Statistical models for the prediction of respirable suspended particulate matter in urban cities. Atmospheric environment, 40(11), 2068-2077.

[4] Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software, 27(3), 1–22.

[5] Liu, T., Lau, A. K., Sandbrink, K., & Fung, J. C. (2018). Time series forecasting of air quality based on regional numerical modeling in Hong Kong. Journal of Geophysical Research: Atmospheres, 123(8), 4175-4196.

[6] Marinov, E., Petrova-Antonova, D., & Malinov, S. (2022). Time Series Forecasting of Air Quality: A Case Study of Sofia City. Atmosphere, 13(5), 788.

[7] Shumway, R. H., & Stoffer, D. S. (2007). Time series analysis and its applications, Fourth Edition. Springer.

[8] World Health Organization. (2021). WHO global air quality guidelines: particulate matter (PM2. 5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization.

[9] Zhang, L., Lin, J., Qiu, R., Hu, X., Zhang, H., Chen, Q., & Wang, J. (2018). Trend analysis and forecast of PM2. 5 in Fuzhou, China using the ARIMA model. Ecological indicators, 95, 702-710.