An Application of ARIMA modelling to air pollution concentrations during covid pandemic in Italy

S Mancini1,a, A Francavilla2, G Graziuso2 and C Guarnaccia2,b

1 Department of Information and Electric Engineering and Applied Mathematics, University of Salerno, via Giovanni Paolo II 132, 84084, Fisciano, SA, Italy
2 Department of Civil Engineering, University of Salerno, via Giovanni Paolo II 132, 84084, Fisciano, SA, Italy

E-mail: asmancini@unisa.it, bceguarnaccia@unisa.it

Abstract. Since the COVID-19 pandemic began, space and ground-based observations have shown how Earth’s atmosphere has observed significant reductions in some air pollutants. Many studies, all over the world, demonstrated how the governmental restrictions imposed because of the spreading of the virus had positive and negative effects on the environment. In this paper, authors discuss how the levels of concentrations of some pollutants varied, in two case studies in Italy, because of the imposed lockdown during the coronavirus pandemic. The extent of the variations CO and PM$_{10}$ has been evaluated by comparing data registered by local monitoring stations, related to the baseline February-May, of three different years, 2018, 2019 and 2020. In order to better assess the variation of the temporal trend of pollutants before (2018, 2019) and during COVID-19 lockdown (2020) proper physic-mathematical models have been applied to the datasets. The calibration and validation of AutoRegressive Integrated Moving Average (ARIMA) models on interesting series of CO and PM$_{10}$ data complete the work.

1. Introduction

Air pollution is one of the most hazardous environmental problem. People health is affected by chemical and physical agents, especially in areas in which high concentrations of pollutants are observed. Thus, it is crucial to develop large and effective monitoring networks, to control the time evolution of the pollutants’ concentrations. The data recorded can be used to calibrate and validate predicting models based on various techniques [1-3]. These models can be used to track sudden changes with respect to the seasonal slope and to implement mitigation actions. In this scenario, an unusual behaviour has been observed during the COVID pandemic of 2020. Several studies highlighted a decrease of some pollutants concentration in urban areas [4]. The most relevant decrease has been observed for NOx. On the contrary, in some regions CO and PM$_{10}$ didn’t show a decrease, due to the relevant increase of heating systems usage and other pollution sources [5].

In this paper, the “Time Series Analysis” (TSA) techniques are applied to a dataset of pollutant concentrations recorded in an Italian city during the first lockdown (spring 2020). These techniques are largely used in several research areas. In [6-10], deterministic decomposition is applied to noise recorded in different zones, produced by various sources, such as traffic noise, airport, etc.. Poisson processes combined with TSA models are also applied to noise in [11]. In [12] respectively the deterministic decomposition and the “Seasonal AutoRegressive Integrated Moving Average” (SARIMA) model of a CO dataset collected in the Metropolitan Area of Monterrey (Mexico) are presented, showing that for a short-term precise prediction SARIMA should be preferred.
After a preliminary statistical description of the dataset, a comparison with average concentrations of previous years will be showed, to discuss possible decrease or increase related to the pandemics. Then, a predictive model will be calibrated and validated on data not used in the hyperparameters and parameters tuning phase.

2. Material and methods

2.1. Time Series Analysis techniques

The Time Series Analysis is the mathematical study of the slope over time of a selected physical variable, measured with a given time basis. When the database includes a single variable measured over time, it can be considered a univariate time series.

The most popular techniques adopted are based on a deterministic decomposition or on Auto-Regressive Integrated Moving-Average (ARIMA) procedures [13].

Generally, when non-stationarity is an issue, the stochastic autoregressive integrated moving average (ARIMA) models are used. They are part of the family of non-stationary linear stochastic processes and, in their most complete version, can be divided into two categories: the ARIMA and the seasonal ARIMA, called SARIMA, used when there is a periodicity in the data series instead. ARIMA predictions evolve over time, using as input recent data, close to the period to be predicted, following the changes of the process. Therefore, the ARIMA models adapt quickly to possible variations of the series, but they pay this quality in terms of short forecast horizon.

The key components of the ARIMA model are the autoregressive or AR component, the differencing component and the moving average or MA component applying the Box-Jenkins method for finding the best fit coefficients. The order of an ARIMA model is usually denoted by the following notation ARIMA(p,d,q), where p, d and q are the order of the autoregressive part, the differencing and the moving-average process, respectively.

The general formula of these models, not including seasonality, can be expressed as following:

\[
\phi_p(B)(1-B)^d Y_t = \theta_q(B) e_t
\]

in which:
- \(Y_t\) is the value of the series observed at the time \(t\),
- \(B\) is the delay operator,
- \(\phi_p\) is the autoregressive polynomials (polynomial AR),
- \(\theta_q\) is the moving average polynomials (polynomial MA),
- \(e_t\) is the residual at the time \(t\), i.e. the difference between the observed and forecasted values at the time \(t\).

Generally, the elaboration work involves four main steps: the dataset analysis and statistical summary, the model calibration, i.e. the tuning of hyperparameters and best ARIMA model selection, the model validation, i.e. the test of the selected model on a dataset not used in the calibration and the model forecasting.

In this paper, the authors will present the calibration of two ARIMA models applied to air pollutants concentrations recorded on a large time range, in two sites described in subsection 2.2. The estimation of the hyperparameters and of the coefficients of the models is done in “\(R\)” software [14], in which also the validation on data not used in the calibration and the forecasting can be run.

2.2. Case studies presentation

The datasets used in this paper have been collected in two fixed monitoring stations installed by the regional agency for environmental protection (Agenzia Regionale per la Protezione dell’Ambiente Campania, ARPAC). The two stations are located in the cities of Nocera Inferiore and Solofra (Fig. 1).
Figure 1. Sites location highlighting the Campania region (in red) and the provinces.

The area of the Solofra station is industrial, with a large presence of tanneries. The tanning, in fact, is a specific expertise of the area, with a large production of leather goods. The production process has several polluting sources related to the refining of leather, with emission of Volatile Organic Compounds (VOCs), Particulate Matter (PM), Hydrogen sulphide (H2S) that is responsible of the typical bad smell of the area. CO and NOx are produced by the thermal plants that continuously provide hot water and steam for the processes. Besides the industrial sources, road traffic, agriculture and other anthropic sources should be considered as well.

The monitoring station of Nocera Inferiore is settled in a strategic urban position to control the changes in emissions mainly related to road traffic, since it is placed in proximity of a highway and close to local roads with large volumes of vehicles. The position of the station is in the residential downtown, with many houses and buildings in the surroundings. Nocera Inferiore is one of the most polluted cities of the Campania Region, being the most polluted city in the province of Salerno. In 2020, for instance, it had 67 thresholds exceedances for PM10 concentrations (50 µg/m³), with respect to the 35 exceedances allowed by the Italian regulation [15].

The two stations record, continuously, Benzene, NOx, SO2, PM10, PM2.5 and CO, together with meteorological data such as temperature and humidity. For the analysis presented in this paper, the authors selected PM10 and CO daily concentrations, recorded from February to May 2020. These datasets have been firstly compared with the same time range in 2018 and 2019, aggregated in monthly levels (Figure 2) and then graphically reported in a time plot (Figure 3) with evidence of the governmental restriction phases.
In particular, the first green vertical line is placed in correspondence of the first publication of national restrictions on the 4th of March 2020 defining the overall closure of schools and the beginning of the first phase of lockdown. The green area ends at the start of phase 2, when a general relaxation of restrictions was implemented.

In particular, the Italian governmental restrictions from the beginning of the pandemics (February 2020), specified by the publication of ministerial decree, can be summarized as in Table 1.

Despite of the decrease of road traffic and production plants operation during the lockdown, it can be noticed that, except for CO concentrations in Solofra and PM\textsubscript{10} concentrations in Nocera Inferiore, the other curves are not drastically different from previous years observations. This is probably due to the compensation of home heating systems increase, since people stayed at home longer than during the same months in 2018 and 2019. In standard conditions, a large part of the population during working days would have been in working places or at schools; these buildings are generally more effective than private houses from the energetic and pollution point of view.

Table 1. Timeline of the Italian governmental restriction from February to May 2020.

| Phase | DPCM\textsuperscript{1} date | Adopted containment Measures |
|-------|-----------------------------|-----------------------------|
| 1     | 4\textsuperscript{th} March 2020 | suspension of all educational activities (all levels schools and universities) |
| 1bis  | 8\textsuperscript{th} March 2020 | total lockdown of all the cities |
| 1ter  | 23\textsuperscript{rd} March 2020 | closure of all activities excepted essential industrial and commercial ones |
| 2     | 17\textsuperscript{th} May 2020 | end of restriction on displacement among cities and regions |

\textsuperscript{1} Decree of the President of the Ministry Council
Figure 3. Average daily levels of PM$_{10}$ and CO registered in Nocera Inferiore and Solofra from February to May 2018, 2019 and 2020 with evidence of the different restriction phases. The red dashed line in PM$_{10}$ plots is the threshold for daily concentration that the Italian regulation allows to overcome 35 days per year.

University of Salerno, for example, currently self-produces about 30% of its energy needs, thanks to photovoltaic roofs and plants, solar thermal, cogeneration and solar cooling [16] providing necessities for more than one thousand persons. Indeed, the University employs about 1000 professors and researchers, about 500 technicians and administrative clerks, plus a large number of PhD students and postdoc fellows, usually working in the labs and in the departments’ rooms, and about 35000 students. All these persons stayed at home, increasing private heating boilers usage, of course. Very often these boilers are based on old functioning systems, with large emissions and gas consumption. Basically, during the lockdown, the pollution sources increased in number, with a worsening in quality.

For the reasons mentioned above, the following analyses will be pursed on CO for the city of Solofra and on PM$_{10}$ datasets for Nocera Inferiore.

2.3. Datasets presentation
As mentioned in the previous subsection, the datasets used in this paper are ‘CO-Solofra’ and ‘PM$_{10}$-Nocera’. The two datasets have been used for the calibration phase.

Each dataset is composed of three datasets of 121 daily concentrations, both for CO and PM$_{10}$ measured in 2018, 2019 and 2020 from the 1st of February to the 31st May. The ‘CO-Solofra’ dataset presented 3 days of missing data in 2020, 2 in 2019 and 3 in 2018. The ‘PM$_{10}$-Nocera’ presented 5 days of missing data in 2020, 8 in 2019 and 9 in 2018. Since time series analysis needs to be performed on continuous dataset missing data has to be imputed. More in details, the missing data (4% of the total calibration data for both pollutants) has been imputed with the mean of the previous and successive measurements, and, for PM$_{10}$, in some cases with the “cold deck” technique, i.e. imputing with the concentrations observed in the same period in the past years. This method of imputation process, did not altered the distribution of the data, preserving the mean and the standard deviation, as showed in Tables 2 and 3.

Table 2. Summary statistics of the CO concentrations measured in Solofra.

| Dataset     | Mean [mg/m$^3$] | Std. Dev. [mg/m$^3$] | Median [mg/m$^3$] | Skew  | Kurt  |
|-------------|-----------------|-----------------------|-------------------|-------|-------|
| Observed    | 0.45            | 0.32                  | 0.33              | 0.71  | -0.77 |
| Reconstructed | 0.45            | 0.32                  | 0.35              | 0.72  | -0.79 |
Table 3. Summary statistics of the PM$_{10}$ concentrations measured in Nocera Inferiore.

| Dataset    | Mean [µg/m$^3$] | Std. Dev. [µg/m$^3$] | Median [µg/m$^3$] | Skew  | Kurt  |
|------------|-----------------|-----------------------|-------------------|-------|-------|
| Observed   | 33.44           | 17.13                 | 29.75             | 0.81  | 0.23  |
| Reconstructed | 33.53           | 17.04                 | 29.50             | 0.80  | 0.19  |

3. Results and discussion

The TSA models, based on ARIMA techniques, calibrated and adopted in this analysis have been selected by using the "auto.arima" function implemented in "R" project software environment minimizing the AIC and BIC criteria, as well as according to the parsimony principle. In particular, the above mentioned automatic function provides hyperparameters and parameters of the best fitting model, using the conditional-sum-of-squares to find starting values, then the maximum likelihood for finding the best values. Then, residual tests need to be performed in order to examine the appropriateness of the model. Thus, the values of the autocorrelation and partial autocorrelation functions (ACF/PACF) of the residual series must all be approximately nil.

For the PM$_{10}$ concentrations in Nocera Inferiore, the best model resulted a simple AR(1), with only order 1 autoregressive component. This choice is suggested by the slope of the autocorrelation and partial autocorrelation functions showed in Figure 4, as well as by the routine "auto.arima" implemented in the forecast package of "R" software.

The graphical comparison between the results of the ARIMA model and the observed data is shown in Figure 5. The overlap between observed PM$_{10}$ concentration and simulation of the model and a one-day delay in the predictions can be easily observed.

![Figure 4. Autocorrelation and Partial autocorrelation for PM$_{10}$ observed in Nocera Inferiore.](image)

![Figure 5. Plot of observed and simulated PM$_{10}$ concentrations in Nocera Inferiore.](image)
The good performance of the model is highlighted also in Figure 6, reporting the bi-sector scatter plot between observations and simulations, and showing that the 80.2% of simulations fall in an interval of one standard deviation from the observations, with some overestimations in the low concentration range and underestimations in the high concentration range.

![Figure 6. Scatterplot of observed and simulated PM$_{10}$ concentrations in Nocera Inferiore](image)

Comparison between observed and simulated data in order to assess the goodness of the selected models has been done, by using similar plots reported in Figures 7, 8 and 9, for CO concentrations in Solofra.

In this case, the series is nonstationary, thus a differentiation has been performed on the data, to work with a smoother time series. The autocorrelation and the partial autocorrelation plots of the differenced series (Figure 7b) suggest a ARIMA(14,1,14) model, that is tested together with the ARIMA(0,1,1) simple model suggested by the BIC criterion (a ranking has been obtained with the “arimald” function of the “ast” package in the “R” software).
Figure 7. Autocorrelation and Partial autocorrelation for CO a) observed series and b) differenced series in Solofra.

Figure 8. Plot of observed and simulated CO concentrations in Solofra, with ARIMA(0,1,1) and ARIMA(14,1,14).

The bisector plot shows that basically all the simulations obtained both with ARIMA(0,1,1) and ARIMA(14,1,14) fall within a one standard deviation interval from the observations, demonstrating that both model have very good performance.
Figure 9. Scatter plot of observed and simulated CO concentrations in Solofra, with ARIMA(0,1,1) on the left, and ARIMA(14,1,14) on the right.

4. Conclusions

Italy has been one of the first European nations to be affected by COVID-19 in 2020 [17]. A wide range of governmental restrictions were adopted from February 2020 to mitigate the pandemic. In this regard, the present research set out to evaluate two aspects: the impact of lockdowns on the concentrations of particulate matter PM$_{10}$ and CO in the Campania region, and how physics-mathematical models, as ARIMA, can perform good simulation of data. Thus, in order to compare short term effects on air quality before and after the coronavirus pandemic two monitoring stations in the Campania region, south Italy, were selected. They are located respectively in the industrial district of Solofra, a famous leather tanning area in the province of Avellino, and in the town of Nocera Inferiore, one of the most densely populated area in the province of Salerno with almost 2200 inhabitants per km$^2$. The aim of this choice was to analyse the effects in two different urban areas: one mainly industrial and another one, Nocera Inferiore, very mixed, characterized by the close presence of residential areas, schools, shops, highways and small industries.

The extent of the variations of the above mentioned pollutants has been evaluated by comparing data registered by local monitoring stations, related to the baseline February-May, of three different years, 2018, 2019 and 2020. As result is evident the limited or insufficient effectiveness of lockdowns in hardly reducing PM$_{10}$ and CO concentrations. Then, by means of ARIMA models, the variation of the temporal trend of pollutants before (2018, 2019) and during COVID-19 lockdown (2020) has been assessed through the calibration and validation of models on interesting selected series: CO for Solofra and PM$_{10}$ for Nocera Inferiore. In particular, the applied ARIMA models showed good performance in the simulation of data.

For further developments, authors will implement the selected models in order to assess the performance in the forecasting phase. Moreover, short and long term effects could be assessed, including also other pollutants, in order to evaluate if and how anti-COVID-19 restrictions influenced Italian citizens’ habits.

References

[1] Cabaneros S, Calautit J K, Hughes B R. 2019. A review of artificial neural network models for ambient air pollution prediction. Env Modelling & Software, 119, pp.285-304, ISSN 1364-8152
[2] Liao K, Huang, X, Dang H, Ren Y, Zuo S, Duan C. 2021. Statistical Approaches for Forecasting Primary Air Pollutants: A Review. Atmosphere, 12, 686.

[3] Guarnaccia C, Lenza TLL, Mastorakis NE and Quartieri J 2011 A comparison between traffic noise experimental data and predictive models results Int J of Mech 5(4), pp. 379-386

[4] Bray CD, Nahas A, Battye WH, Aneja VP. Impact of lockdown during the COVID-19 outbreak on multi-scale air quality. 2021. Atmos Environ (1994), 254, 11836.

[5] Thieriot H., Myllyvirta L. Air pollution returns to European capitals: Paris faces largest rebound.2020. Centre for Research on Energy and Clean Air (CREA)

[6] Guarnaccia C, Bandeira J, Coelho MC, Fernandes P, Teixeira J, Ioannidis G and Quartieri J 2018 Statistical and semi-dynamical road traffic noise models comparison with field measurements AIP Conf. Proc. 1982 pp1-6

[7] Guarnaccia C, Quartieri J, Tepedino C. 2017. Deterministic decomposition and seasonal ARIMA time series models applied to airport noise forecasting. Proc. of the Int. Conf. on Applied Mathematics and Computer Science, AIP Conference Proceedings 1836, pp. 1-7.

[8] Zhang X, Zhao M, Dong R 2020. Time-Series Prediction of Environmental Noise for Urban IoT Based on Long Short-Term Memory Recurrent Neural Network Appl. Sci. 10, 1144, 1-18

[9] Guarnaccia C, Quartieri J, Rodrigues ER and Tepedino C 2014 Acoustical noise analysis and prediction by means of multiple seasonality time series model, Int. J. Math. Mod. Meth. Appl. Sci. 8, 384-393

[10] Steele C 2001 A critical review of some traffic noise prediction models Appl. Acoust. 62 pp 271

[11] Guarnaccia C, Quartieri J, Tepedino C, Rodrigues E R. 2016. A time series analysis and a non-homogeneous Poisson model with multiple change-points applied to acoustic data. Appl. Acou, 114, pp. 203–212.

[12] Guarnaccia C, Cerón Bretón J G, Quartieri J, Tepedino C, Cerón Bretón C R. 2014. An Application of Time Series Analysis for Forecasting and Control of Carbon Monoxide Concentrations, Int. J. of Math. Models and Methods in Applied Sciences, 8, pp 505-515.

[13] Cryer P D, Chan K. Time Series Analysis, with applications in R. 2008.Second Edition, Springer.

[14] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

[15] Gazzetta Ufficiale della Repubblica Italiana (Italian Republic) 2010. Ministerial Decree August 13, 2010 n.155. Available online at: https://www.camera.it/parlam/leggi/deleghe/testi/10155dl.html (last access July 20, 2021)

[16] https://web.unisa.it/en/campus-life/campus/photovoltaic-park (last access July 20, 2021)

[17] Rovetta A. 2021. The Impact of COVID-19 Lockdowns on Particulate Matter Emissions in Lombardy and Italian Citizens’ Consumption Habits. Front. Sustain. 2:649715.

Acknowledgments
Authors would like to acknowledge Marialuisa Ciccone for contributing to this research during her bachelor thesis.