DOI: 10.26314/GARR-Conf22-proceedings-02

A real-world case study of time series analysis for air quality control

Maria Laura Clemente

CRS4, Italy

Abstract. The availability of data from the web enables interesting discrete time series analysis in terms of trends and in order to look for correlations between more aspects. An in-progress research activity about time series analysis with machine learning applied to air pollutants is presented, based on real data collected for the Cagliari Metropolitan area; correlations with weather conditions have been explored as well. For the research activity the ARIMA algorithms have been applied, which are made of the combinations of three different models: Auto Regressive (AR), Integrated (I), and Moving Average (MA). First results have been obtained which require further investigation.

Keywords. Time Series Analysis, Air Quality, ARIMA algorithms

Introduction

Time series are datasets providing the values of a variable over time. The availability of time series data from the web enables interesting analysis in terms of trends, correlations, and predictions. Time series analysis is made of three fundamental parts (shown in Figure 1):

- descriptive analysis, which is related to past values,
- predictive analysis, which starting from past values, try to predict future ones, with some degree of probability,
- prescriptive analysis, which is based on the predictions to plan possible actions.

In practice, time series analysis is useful to translate sequences of historical data into valuable information for decision makers in strategic plan.

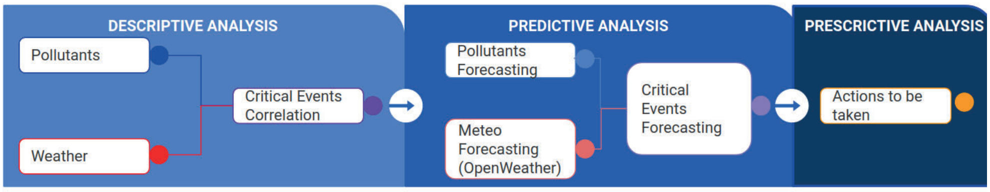
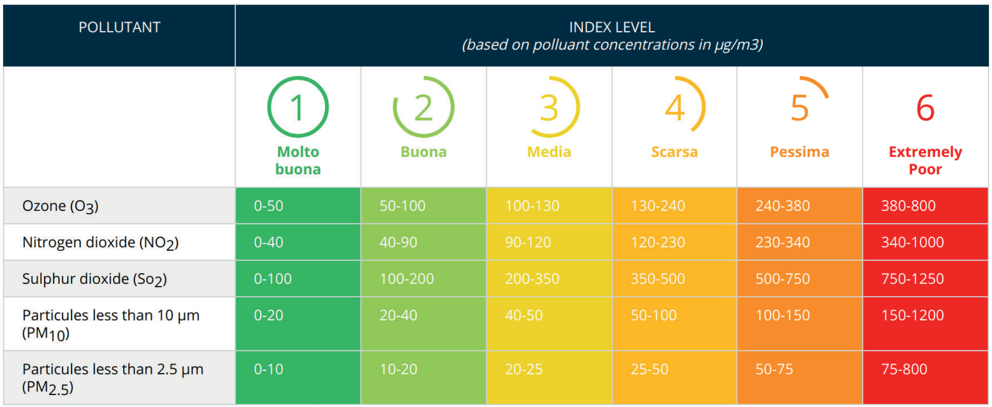


Fig. 1 Steps of time series analysis

There are many works about time series analysis, some of them are particularly practical providing clear explanations and are available online, such as, Hyndman et al. 2018, Brockwell et al. 2016, Peña et al. 2000.

In literature many studies are related to the control of pollutants in the air, their trends

and their correlations with weather conditions, using ARIMA models: for the Pune Region, (Sudumbrekar et al. 2021), for the city of Chennai (Mani et al. 2021), for the city of Hyderabad (Gopu et al. 2021), for the Port of Igoumenitsa (Spyrou et al. 2022), and for the city of Vancouver (see link to kaggle Portal in References), just to mention some examples. Other studies analysed correlations of weather conditions, pollutant concentration in the air and mortality (Gouveia et al. 2000) or diseases of the respiratory system (Rossi et al. 1993, Li et al. 2022, Xiong et al. 2022).
In general, the pollutants considered for these studies are Ozone (O3), Nitrogen dioxide (NO2), Sulphur dioxide (SO2), Particles less than 10 µm (PM10), Particulate less than 2.5 µm (PM2.5), benzene (C6H6), and Carbon Monoxide (CO). These types of studies typically refer to the Air Quality Index (AQI) which is based on the first 5 pollutants in the previous list. Unfortunately, there isn't a unique worldwide standard for it, with a fixed number of classes and thresholds. The European Air Quality Index (shown in Figure 2) is managed and calculated for many locations by the Copernicus Atmosphere Monitoring Service (see link to Copernicus in References) and this European scale has been adopted also for this study (more details in the Handbook about the European Air Quality legislation, the European Parliament Directive 2008/50/EC, and the related Italian D.Lgs.



POLLUTANT	INDEX LEVEL *(based on polluant concentrations in µg/m3)*					
	1 Molto buona	2 Buona	3 Media	4 Scarsa	5 Pessima	6 Extremely Poor
Ozone (O_3)	0-50	50-100	100-130	130-240	240-380	380-800
Nitrogen dioxide (NO_2)	0-40	40-90	90-120	120-230	230-340	340-1000
Sulphur dioxide (So_2)	0-100	100-200	200-350	350-500	500-750	750-1250
Particules less than 10 µm (PM_{10})	0-20	20-40	40-50	50-100	100-150	150-1200
Particules less than 2.5 µm ($PM_{2.5}$)	0-10	10-20	20-25	25-50	50-75	75-800

Nota: I valori di PM10 e PM2.5 si basano su una media mobile a 24 ore

Fig. 2
The European
Air Quality Index

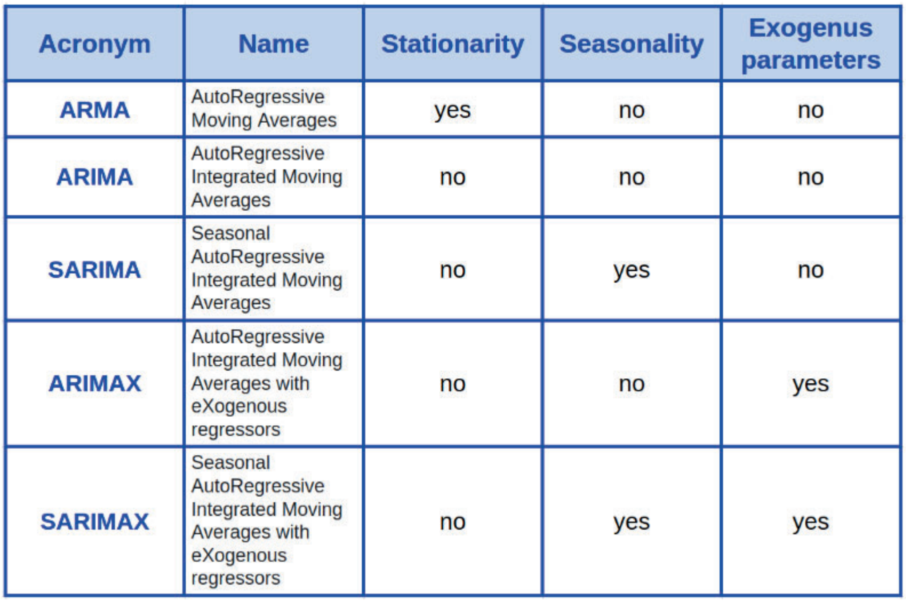
155/2010, in the References)
It's important to note that the daily value of the EU AQI is computed as the highest values of the five pollutants. So, for instance, if during a day the average value of ozone was inside the range level 3 (100-130 µm/m3), while all the other four values are inside the ranges of lower levels, the EU AQI for that day will result equal to 3.

1. Methodology

The activity was carried out using python scripts importing many well-known open-source libraries such as: numpy, pandas, matplotlib, seaborn, missingno, statsmodels, scipy, sklearn, and pmdarima. Considering the above mentioned three main steps of time series analysis, the presented activity must be considered a work in progress because it has covered topics belonging to the first two, as described here after.

During the Descriptive Analysis, the time series have been collected: pollutants (source ARPA Sardegna) and weather conditions (source Open Weather). From the first analysis of the available data, some gaps in the historical series where observed, which required a pre-processing through interpolation. After the pre-processing and analysis of the dataset, the daily EU AQI has been computed. Then a graph of the time series for each pollutant and for each detection station has been analysed and statistical metrics have been computed, trends, seasonality and irregularities were looked for, along with the level of stationarity. A value of linear correlation between all the time series has been analysed, and non-linear correlations were looked for through scatter plots of two time-series at a time.

The Predictive Analysis started with the dataset split in train set (80%) and test set (20%) as preparation for the ARIMA algorithms (summarized in Figure 3), made of a combination of: Auto Regressive (AR), Integrated (I), and Moving Average (MA).



Acronym	Name	Stationarity	Seasonality	Exogenus parameters
ARMA	AutoRegressive Moving Averages	yes	no	no
ARIMA	AutoRegressive Integrated Moving Averages	no	no	no
SARIMA	Seasonal AutoRegressive Integrated Moving Averages	no	yes	no
ARIMAX	AutoRegressive Integrated Moving Averages with eXogenous regressors	no	no	yes
SARIMAX	Seasonal AutoRegressive Integrated Moving Averages with eXogenous regressors	no	yes	yes

Fig. 3
Scheme of ARIMA algorithms

The aim is to find out the best forecasting with the optimal combination of the parameters p, d, and q (Hannan 1980). The model is initialized and trained with the training set; for each result, the Akaike Information Criteria (AIC, Akaike, 1978) and the Bayesian Information Criteria (BIC) are used to choose the best parameters for the model, which is the combination producing the minimum values of AIC and BIC (Hyndman et al. 2018). After this training, the best combination of parameters is used to run the model on the test set. Then, the predictions obtained for the test set are compared to the real values (which for the test set are known) and the error can be evaluated, in terms of Mean Absolute Error. These values are then used to compare all the ARIMA models (depending on the combinations of parameters p, d, and q).

3. Results

Although the presented activity is still in progress, and a proper forecasting wasn't validated yet, the results obtained so far allow to draw some considerations.

In all the monitoring stations of the study area, during the period of observation, an EU_AQI of level 2 (corresponding to a qualitative score 'good') resulted the more frequent; in some of the locations there have been different situations, for instance, in one of them the values of EU_AQI presented less than 150 days of level 1 and a comparable number of days of level 3 and 4 (shown in Figure 4).

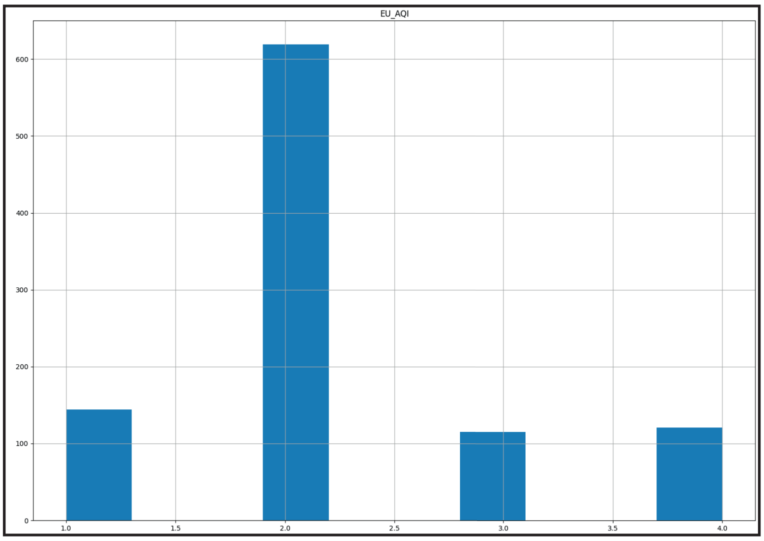


Fig. 4
Values of EU_AQI at one of the monitoring stations in the study area

Interesting correlations were found, for instance, between ozone and average temperature and inverted correlations between ozone and humidity, confirmed by the scattered plots shown in Figure 5, related to the values measured in one of the monitoring stations of the study area. As expected, ozone concentration in the air increases during the summer because it is formed by the interaction of sunlight with hydrocarbons and nitrogen oxides emitted by traffic and industries like refineries. For the same reason, to this ozone higher values corresponds the minimum in nitrogen dioxide (as evident in Figure 6).

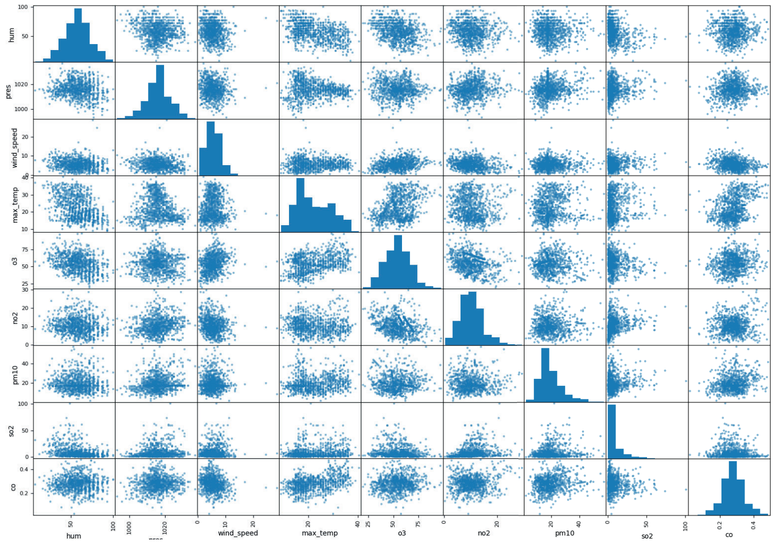


Fig. 5
Correlation Matrix between pollutants and weather conditions

As expected, cyclic behaviours could be evident in the time series, although it wouldn't be correct calling them 'seasonal' because pollutants and weather conditions don't repeat the same peaks and troughs and the cycles change in length from year to year.

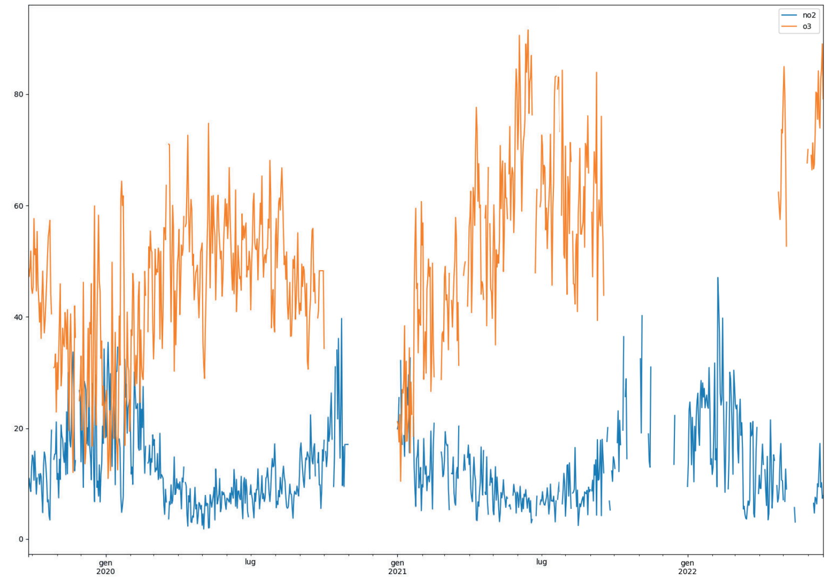


Fig. 6
Ozone and Nitrogen
Dioxide cyclic plotting

In other words, they are stationary, as confirmed by the Augmented Dickey-Fuller (ADF) test which was used: the p-value was considered, which brings to a non-stationary hypothesis for values greater than 0.05, and stationary for values equal or less than 0.05. So, the Integrated part of the ARIMA model could be reduced to ARMA (because the parameter d could be set always to zero). Some first predictions could be elaborated, as the one shown in Figure 7, which is related to CO values measured at one of the monitoring stations of the study area.

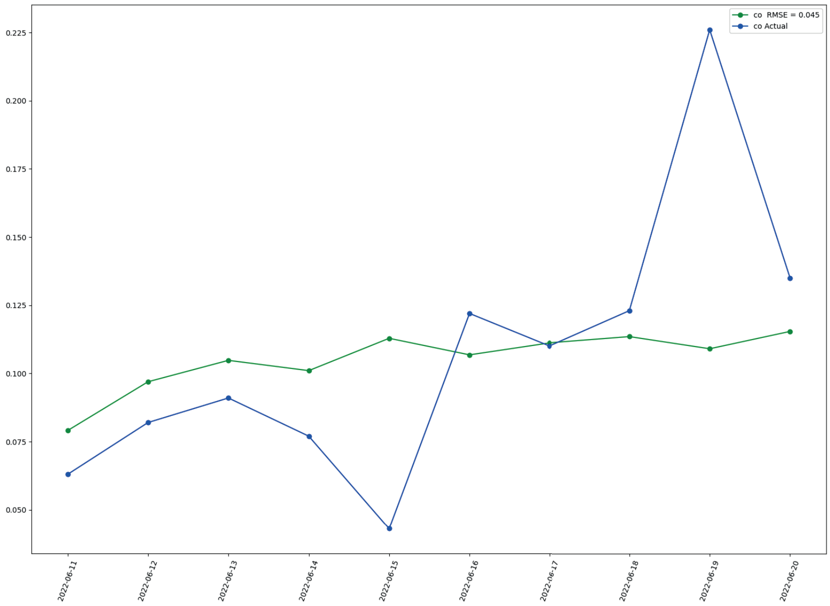


Fig. 7
Prediction of CO values
obtained with ARIMA

4. Conclusions and Future Work

A work in progress activity about air quality analysis in the Cagliari Metropolitan Area was presented, although more data shall be collected, and further scientific investigations shall be carried out through ARIMA models.

A future interesting work could be conducted to improve the study through a correlation analysis between weather condition and air quality with incidence of respiratory diseases.

Acknowledgments

This work was partially supported by the Sardinian Local Government (RAS), within the Joint Innovation Center (JIC) project.

A special thanks to Carlo Impagliazzo for providing the data used for this research activity.

References

Akaike, H., 1978. Time series analysis and control through parametric models. In D. F. Findley (Ed.), Applied time series analysis. New York: Academic.

Bedekar G.B., Patil R.S., Tergundi P., Goudar R. H. (2021), An Efficient Implementation of ARIMA Technique for Air Quality Prediction. Available at SSRN: https://ssrn.com/abstract=3889537 or http://dx.doi.org/10.2139/ssrn.3889537

Brockwell, P.J., Davis, R.A. (2016), Modelling and Forecasting with ARMA Processes. In: Introduction to Time Series and Forecasting. Springer Texts in Statistics. Springer, Cham. https://doi.org/10.1007/978-3-319-29854-2_5

Copernicus Atmosphere Monitoring Service CAMS https://www.euronews.com/weather/copernicus-air-quality-index

European Parliament 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; Official Journal of the European Union, Europe: 2008.

Gopu P., Panda R.R., Nagwani N.K. (2021). Time Series Analysis Using ARIMA Model for Air Pollution Prediction in Hyderabad City of India. In: Reddy, V.S., Prasad, V.K., Wang, J., Reddy, K.T.V. (eds) Soft Computing and Signal Processing. Advances in Intelligent Systems and Computing, vol 1325. Springer, Singapore. https://doi.org/10.1007/978-981-33-6912-2_5

Gouveia N., Fletcher T. (2000), Time series analysis of air pollution and mortality: effects by cause, age and socioeconomic status. Epidemiol Community Health; 54, pp. 750–755.

Handbook about European Air Quality Legislation
https://ec.europa.eu/environment/archives/enlarg/handbook/air.pdf

Hannan, E. J., 1980. The Estimation of the Order of an ARMA Process." The Annals of Statistics, vol. 8, no. 5, 1980, pp. 1071–81. JSTOR, http://www.jstor.org/stable/2240437.

Hssain S., Ahmed S., Uddin J., (2021), Impact of weather on COVID-19 transmission in south Asian countries: An application of the ARIMAX model, Science of The Total Environment, Volume 761, 20 March 2021, 143315

Hyndman, R.J., Athanasopoulos, G. (2018), Forecasting: principles and practice, 3nd edition, OTexts: Melbourne, Australia. https://otexts.com/fpp3/ accessed on 04 July 2022.

Kaggle portal, 2020
Vancouver Air Pollution: Time Series Analysis

https://www.kaggle.com/lucyhu/vancouver-air-pollution-time-series-analysis
Li H., Ge M., Zhang, M. (2022), Spatio-temporal distribution of tuberculosis and the effects of environmental factors in China. BMC Infect Dis 22, 565. https://doi.org/10.1186/s12879-022-07539-4
Mani G., Viswanadhapalli J.K., Stonier A.A. (2021), Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models. Journal of Engg. Research Vol.10 No. (2A) pp. 179-194. DOI: 10.36909/jer.10253
Peña D., Tiao G.C., Tsay R.S. (2000), A course in time series analysis. John Wiley & Sons.
Repubblica Italiana, D.Lgs. 155/2010-Attuazione Della Direttiva 2008/50/CE Relativa Alla Qualità Dell'aria Ambiente e per Un'aria più
Pulita in Europa; Repubblica Italiana: Rome, Italy, 2010.
Rossi O.V., Kinnula V.L., Tienari J., Huhti E. (1993). Association of severe asthma attacks with weather, pollen, and air pollutants.
http://dx.doi.org/10.1136/thx.48.3.244
Spyrou E.D., Tsoulos I., Stylios C. (2022), Applying and Comparing LSTM and ARIMA to Predict CO Levels for a Time-Series Measurements in a Port Area.
Signals 2022, 3(2), 235-248; https://doi.org/10.3390/signals3020015
Sudumbrekar A., Kale R., Kaurwa T., Mule V., and Devkar A. (2021), Study of ARIMA Model for PM2.5 Prediction using Real-World Data Gathered from Pune Region. In book: New Frontiers in Communication and Intelligent Systems, pp.105-111, 2021.
DOI:10.52458/978-81-95502-00-4-13
Xiong Y., Yang M., Wang Z., Jiang H., Xu N., Tong Y., Yin J., Chen Y., Jiang Q., Zhou Y. (2022). Association of Daily Exposure to Air Pollutants with the Risk of Tuberculosis in Xuhui District of Shanghai, China. Int. J. Environ. Res. Public Health 2022, 19, 6085. https://doi.org/10.3390/ijerph19106085

Author

Maria Laura Clemente clem@crs4.it

Maria Laura Clemente, senior technologist, holds a master's degree in Civil Engineering (1994) and has been working at CRS4 since 1996. She has been involved in research and development activities using many programming languages, such as Python, Javascript, C, C++ and Java. She has developed microscopic traffic simulations with SUMO (sumo.dlr.de). She is experienced in Artificial Intelligent algorithms for content personalization, recommender systems, object detection and tracking, time series analysis and prediction. She has served as expert for the European Commission in the review of H2020 project proposals.