ECAS 2020

sciforum

Conference Proceedings Paper

Application of a machine learning methodology for data implementation

Chris G. Tzanis *, Anastasios Alimissis and Ioannis Koutsogiannis

Climate and Climatic Change Group, Section of Environmental Physics and Meteorology, Department of Physics, National and Kapodistrian University of Athens, Greece
chtzanis@phys.uoa.gr (C.G.T.), alimiss@phys.uoa.gr (A.A.), koutsog@phys.uoa.gr (I.K.)
* Correspondence: chtzanis@phys.uoa.gr

Received: date; Accepted: date; Published: date

Abstract: An important aspect in environmental sciences is the study of air quality, using statistical methods (environmental statistics) which utilize large datasets of climatic parameters. The air quality monitoring networks that operate in urban areas provide data of the most important pollutants, which via environmental statistics can be used for the development of continuous surfaces of pollutants concentrations. Generating ambient air quality maps can help guide policy makers and researchers to formulate measures to minimize the adverse effects. The information needed for a mapping application can be obtained by employing spatial interpolation methods to the available data, for generating estimations of air quality distributions. This study uses point monitoring data from the network of stations that operates in Athens. A machine learning scheme will be applied as a method to spatially estimate pollutants' concentrations and the results can be effectively used to implement missing values and provide representative data for statistical analyses purposes.

Keywords: artificial neural networks; shallow neural networks; machine learning; spatial interpolation; data implementation; air quality

1. Introduction

Studying the distribution of air quality parameters is an important task of urban communities. According to the European Environmental Agency (EEA), air pollution is identified as a major environmental health hazard in Europe as hundreds of thousands of Europeans are affected each year by air quality issues [1-3]. Effective planning strategies require constant monitoring of the various pollutants, creating databases suitable for statistical analysis. Increased data availability can help researchers produce more reliable results. Spatial interpolation techniques have been widely used in air quality studies [4-5] as they can be utilized for data implementation in pollutant time series with missing values and even for sites of interest with no data availability. Additionally, by using these implemented databases, the development of informational tools such as Air Quality Indices (AQI) can be beneficial for presenting in a comprehensible manner new insight to policy makers and the public [6-8]. The EEA proposed a European Air Quality Index (EAQI) which is based on hourly concentrations of five key pollutants (PM_{10}, $PM_{2.5}$, NO_2, O_3 and SO_2) and has six different levels based on each pollutant's concentrations. This study aims to present a methodology for filling gaps in environmental sciences and specifically in the field of air quality. From the original datasets and based on concentration time series for the selected pollutants of the EAQI, a machine learning data implementation process was followed. This methodology can be utilized as a fast and effective

tool which will contribute to the development of indexes such as the EAQI which will subsequently visualize air pollutants' profiles and provide insight in patterns and relationships.

2. Experiments

2.1 Data

The air quality monitoring sites, from which the data were derived, are located at the metropolitan city of Athens, in Greece. As part of the southeastern Mediterranean region, Athens climate is defined by dry summers (long periods, during which the temperatures are considerably high) and wet winters (these periods are usually short) [9]. The basin is bounded by mounts Parnitha, Pentelikon, Hymmetus and Aigaleo to the north, northeast, east-central and west respectively. Due to the transport mechanisms, the topography of the area and the proximity to the sea, the air pollution fields are greatly affected by various flows of different scales [10-13]. The monitoring sites in the area, are part of an air quality monitoring network that operates since 1984, under supervision of the Hellenic Ministry of Environment and Energy (MEE). The network is considered representative of the pollutants' spatial variability and thus suitable for the application of advanced statistical methodologies. For the development of the EAQI, a different number of stations was selected for each pollutant. The criterion for this selection was that a station should have at least a small percent of available data and thus, could contribute to the data implementation methodology. For the five pollutants, NO_2, O_3, PM_{10}, $PM_{2.5}$ and SO_2, the number of stations used was fourteen, thirteen, eleven, six and six, respectively. All five were monitored hourly, and the time period of the analysis was three years (2016-2018).

2.2 Methodology

The first step in this study, after the database development, was to find the number of gaps that are present in each station's data (target station/missing hourly concentrations) for 2018. This task was performed for all pollutants individually. However, in order to be able to apply effectively the machine learning spatial interpolation scheme, a specific criterion was adopted. For each one of these gaps at a target station, at the same time all the remaining stations must have an available measurement. Even if one of them had also a gap, it was not included in the interpolation process. The results of this step are presented in Table 1 and reveal the number of missing values that can be potentially estimated and used to increase the available data points. The next step was to apply an Artificial Neural Network (ANN) approach for spatial estimation purposes. To achieve this, a Shallow Neural Network (SNN) was utilized as a practical and fairly simple ANN that is moderately demanding in terms of time and computational power. However, it can effectively simulate complex nonlinear relationships between parameters. In detail, two-layer networks with sigmoid hidden neurons and linear output neurons were used (Figure 1).

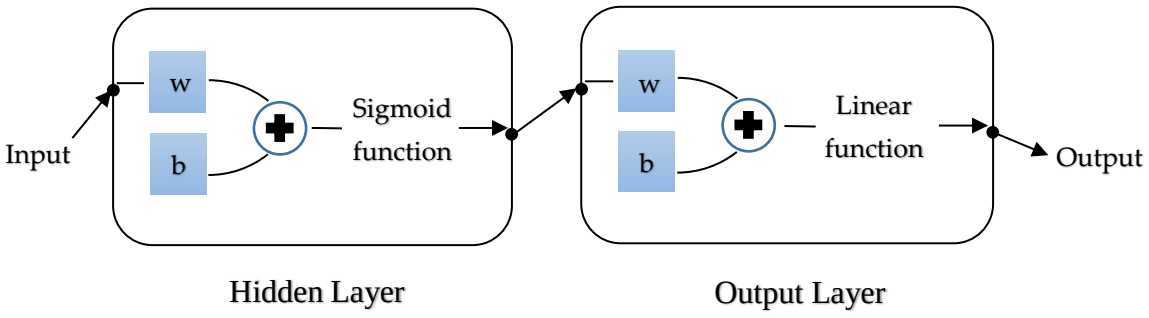

Figure 1. A two-layer network with sigmoid hidden neurons and linear output neurons

The number of hourly concentrations that were used for the models were those for which none of the stations had a missing value. The training of the networks was performed with the Levenberg-Marquardt backpropagation algorithm. The dataset was divided into three subsets used for training, validation and testing randomly and each subset corresponded to specific percentages of the original data (70% training, 15% validation, 15% testing). Depending on the pollutant, the number of data points used for the subsets was different and is presented in Table 2. The network architecture includes a number of inputs equal to the number of all stations minus the target station (13 for NO_2, 12 for O_3, 10 for PM_{10}, 5 for $PM_{2.5}$ and 5 for SO_2), while the output is always one (target station). Regarding the number of neurons in the hidden layer, the performance of each network was evaluated by using the Mean Absolute Error (MAE) statistical criterion [14-18], which is calculated by using the following equation:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|E_i - O_i| \quad (1)$$

where E denotes the estimated concentration, O the observed concentration and n the number of data points. Lower MAE values illustrate the optimum performing network. Five runs were performed for all schemes and for hidden layer neurons that ranged from 1 to 40. The best performing networks and their architecture are presented in Table 3. By using these selected SNN models for the corresponding inputs of 2018, the gaps in each station and pollutant were filled.

3. Results

A total of 12,526 missing values were estimated and the percentage of gaps that were filled out in each station was above 40% for PM_{10} and $PM_{2.5}$, above 20% for O_3 and SO_2 and above 15% for NO_2. Regarding O_3 and NO_2 where the percentage of interpolated values is lower, it needs to be considered that they had a higher number of input stations and thus, the criterion that none of the inputs should have a missing value for each gap of the target station, was more difficult to fulfill. Table 1 presents in detail the gaps originally and after the interpolation, as well as the percentage of missing values that were estimated.

Table 1. Number of missing values (gaps) during 2018, for the original and spatially interpolated dataset

	Original gaps	Gaps after interpolation	Difference	Estimated percentage (%)
NO_2	13,253	11,145	2,108	15.91
O_3	10,814	7,961	2,853	26.38
PM_{10}	7,182	3,948	3,234	45.03
$PM_{2.5}$	4,558	2,524	2,034	44.62
SO_2	7,043	4,746	2,297	32.61

The number of data points for the training, validation and testing subsets and for each pollutant are presented in Table 2. Pollutants with lower number of input stations are associated with higher data points numbers per station (smaller probability for all the stations to have a missing value at the same time). However, more stations (NO_2, O_3) provide additional data points. NO_2 and $PM_{2.5}$ are the pollutants which provided more data for training, validation and testing purposes.

Table 2. Number of data points distributed to the training, validation and testing subset for the 2016-2017 time period

	Training	Validation	Testing	Total
NO_2	47,151	10,101	10,101	67,353
O_3	25,272	5,412	5,412	36,096
PM_{10}	13,410	2,880	2,880	19,170
$PM_{2.5}$	37,785	8,100	8,100	53,985
SO_2	13,925	3,080	3,080	20,085

The architecture of the optimum performance models is presented in Table 3. The hidden neurons number is an average of all the stations for each pollutant. The MAE average values (measured in the same units as the concentrations of the pollutants, $\mu g/m^3$) in these cases are also included. However, all pollutant-specific networks have the same number of inputs and all networks have a single output (the target station). The average hidden neuron value ranges from 21.7 to 25.2 which reveals that the models are at an almost equal complexity level.

Table 3. Number of input, hidden (average) and output neurons as well as MAE (average), mean concentration values and percentage of error (MAE to mean concentration) for the best performing models and the 2016-2017 time period

	Input neurons	Hidden	Output	MAE	Mean	Error (%)
NO_2	13	21.7	1	5.80	32.70	17.74
O_3	12	22.3	1	6,86	58.86	11.65
PM_{10}	10	23.6	1	5.71	29.53	19.34
$PM_{2.5}$	5	25.2	1	5.17	23.81	21.71
SO_2	5	22.5	1	1.89	6.06	31.19

4. Discussion

According to Table 3 results, it can be concluded that the error percentage is higher when the number of input stations is lower and subsequently the information provided for training is more limited. O_3 is an exception to this statement because although the number of input stations is 12 versus 13 for NO_2 and correspondingly the available data points are nearly half, the error percentage is considerably lower. This can be explained by examining other behavioral characteristics of this pollutant (differences in mean values among stations, more easily identifiable patterns in datasets etc.). When comparing $PM_{2.5}$ and SO_2, where the input neurons are five for both, the prediction performance for SO_2 is lower, possibly due to the smaller number of data points, according to Table 2 ($PM_{2.5}$ has nearly three times more data points). Different approaches to evaluate the performance of the models can be followed (scatter diagrams, correlation metrics, etc.) as well as more types of similar complexity neural network models can be examined.

5. Conclusions

This study applied SNN models as a tool for point spatial interpolation of air quality parameters, using data from an air quality monitoring network located at a densely populated urban area. Five air quality parameters were selected (PM_{10}, $PM_{2.5}$, NO_2, O_3 and SO_2), due to their importance in the field of air quality indexes and more specifically, based on the EAQI (proposed by EEA). The results highlight that the models' performance is significantly affected by the density of the air quality monitoring network (number of stations and data points per station) as well as the specific patterns that characterize each pollutant's concentrations. The training dataset is crucial for the networks' development and needs to be carefully selected in order to provide adequate information which will augment the networks' generalization ability. This work can be utilized as an alternative for

commonly used spatial interpolation methods in the field of air quality and further improvements can be made by using more advanced networks and/or adding meteorological parameters as inputs.

Author Contributions: C.G.T. and A.A. were involved into the conceptualization, writing-original draft preparation and writing-review and editing of this work, while individually C.G.T. was responsible for the data curation, validation of the results and supervised the whole procedure. All authors performed the various steps of the methodology, processed the data and developed the neural network models. All authors were involved in the discussion of the results and commented on the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to acknowledge the Ministry of Environment and Energy for providing the air quality parameters' database which was utilized in this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Amanollahi, J., Tzanis, C., Abdullah, A. M., Ramli, M. F. & Pirasteh, S. (2013). Development of the models to estimate particulate matter from thermal infrared band of Landsat Enhanced Thematic Mapper. *International Journal of Environmental Science and Technology*, 10 (6), 1245–1254.
2. Baklanov, A., Molina, L. T., & Gauss, M. (2016). Megacities, air quality and climate. *Atmospheric Environment*, *126*, 235–249. https://doi.org/10.1016/j.atmosenv.2015.11.059.
3. European Environment Agency. (2013). Air quality in Europe—2013 Report: EEA report no 9/2013. In European Union. Retrieved from http://www.eea.europa.eu/publications/air-quality-in-europe-2013.
4. Can, A., Dekoninck, L., & Botteldooren, D. (2014). Measurement network for urban noise assessment: Comparison of mobile measurements and spatial interpolation approaches. *Applied Acoustics*, 83, 32–39. https://doi.org/10.1016/j.apacoust.2014.03.012.
5. Denby, B., Sundvor, I., Cassiani, M., de Smet, P., de Leeuw, F., & Horálek, J. (2010). Spatial Mapping of Ozone and SO2 Trends in Europe. *Science of the Total Environment*, 408(20), 4795–4806. https://doi.org/10.1016/j.scitotenv.2010.06.021.
6. Zhan, D., Kwan, M. P., Zhang, W., Yu, X., Meng, B., & Liu, Q. (2018). The driving factors of air quality index in China. *Journal of Cleaner Production*, 197, 1342–1351. https://doi.org/10.1016/j.jclepro.2018.06.108.
7. Silva, L. T., & Mendes, J. F. G. (2012). City Noise-Air: An environmental quality index for cities. *Sustainable Cities and Society*, 4(1), 1–11. https://doi.org/10.1016/j.scs.2012.03.001.
8. Ganguly, N. D., Tzanis, C. G., Philippopoulos, K. & Deligiorgi, D. (2019). Analysis of a severe air pollution episode in India during Diwali festival – a nationwide approach. *Atmósfera*, 32(3), 225–236.
9. Tzanis, C. G., Koutsogiannis, I., Philippopoulos, K. & Deligiorgi, D. (2019). Recent climate trends over Greece. *Atmospheric Research*, 230, 104623, doi:10.1016/j.atmosres.2019.104623.
10. Tzanis, C. G., Alimissis, A., Philippopoulos, K., & Deligiorgi, D. (2019). Applying linear and nonlinear models for the estimation of particulate matter variability. *Environmental Pollution*, 246, 89–98. https://doi.org/10.1016/j.envpol.2018.11.080.
11. Deligiorgi, D., Philippopoulos, K., Thanou, L., & Karvounis, G. (2009). A Comparative Study of Three Spatial Interpolation Methodologies for the Analysis of Air Pollution Concentrations in Athens, Greece. *AIP Conference Proceedings*, 1203, 445–450.
12. Tzanis, C. & Varotsos, C. A. (2008). Tropospheric aerosol forcing of climate: a case study for the greater area of Greece. *International Journal of Remote Sensing*, 29 (9), 2507–2517.
13. Varotsos, C., Christodoulakis, J., Tzanis, C. & Cracknell, A. P. (2014). Signature of tropospheric ozone and nitrogen dioxide from space: A case study for Athens, Greece. *Atmospheric Environment*, 89, 721–730.
14. Cort, J. W., & Kenji, M. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79–82. https://doi.org/10.3354/cr00799.
15. Alimissis, A., Philippopoulos, K., Tzanis, C. G., & Deligiorgi, D. (2018). Spatial estimation of urban air pollution with the use of artificial neural network models. *Atmospheric Environment*, 191, 205–213. https://doi.org/10.1016/j.atmosenv.2018.07.058.

16. Fallahi, S., Amanollahi, J., Tzanis, C. G., & Ramli, M. F. (2018). Estimating solar radiation using NOAA/AVHRR and ground measurement data. *Atmospheric Research*, 199(September 2017), 93–102. https://doi.org/10.1016/j.atmosres.2017.09.006.
17. Rahimpour, A., Amanollahi, J. & Tzanis, C. G. (2020). Air quality data series estimation based on machine learning approaches for urban environments. *Air Quality, Atmosphere & Health*, doi:https://doi.org/10.1007/s11869-020-00925-4.
18. Mirzaei, M., Amanollahi, J. & Tzanis, C. G. (2019). Evaluation of linear, nonlinear, and hybrid models for predicting $PM_{2.5}$ based on a GTWR model and MODIS AOD data. *Air Quality, Atmosphere & Health*, 12 (10), 1215–1224.