atmosphere

MDPI

Article

Addressing Missing Environmental Data via a Machine Learning Scheme

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, 15784 Athens, Greece; alimiss@phys.uoa.gr (A.A.); koutsog@phys.uoa.gr (I.K.)

* Correspondence: chtzanis@phys.uoa.gr

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 on 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 used point-monitoring data from the network of stations that operates in Athens, Greece. A machine-learning scheme was 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; missing data; air quality

Citation: Tzanis, C.G.; Alimissis, A.; Koutsogiannis, I. Addressing Missing Environmental Data via a Machine Learning Scheme. *Atmosphere* **2021**, *12*, 499. https://doi.org/10.3390/atmos12040499

Academic Editor: Jia Xing

Received: 18 February 2021
Accepted: 10 April 2021
Published: 15 April 2021

1. Introduction

Studying the distribution of air-quality parameters is an important task of urban communities. According to the European Environment 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]. Furthermore, air-quality parameters' concentrations are associated with effects that are non-health-related and can influence the interactions between humans and the environment that surrounds them [4,5]. 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. However, for areas where the number of air-quality-monitoring sites that are part of a network is limited and/or not fully functional (possibly due to high establishment and maintenance costs, etc.) and, subsequently, a lower number of available observations cannot reflect the spatiotemporal distribution; thus, interpolation methods are of great significance. Spatial interpolation techniques have been widely used in air-quality studies [6,7], as they can be utilized effectively for data implementation in pollutant time series with missing values and even for sites of interest with no data availability. There are several categories in which these techniques can be classified. According to Li and Heap [8], they can be typically grouped into non-geostatistical, geostatistical and combined methods. The importance of using these methodologies to fill data gaps has been proved by the abundance of research studies on this subject, which, especially during the last few years, emphasize the need for the development of more advanced methodologies [9–12]. Machine learning (ML) and, in particular, artificial neural networks (ANNs) are considered as a novel superior alternative to traditional data implementation techniques, due to their ability to perceive the relationships among the various air-quality parameters as

nonlinear in contrast with other statistical schemes which assume that these linkages are linear [13,14]. While ANNs have been mostly utilized for temporal predictions in the field of air-quality and climatic-parameters forecasting [15–18], they have been additionally applied as a tool to provide spatial estimations in order to create datasets without missing values [12,19–21]. Additionally, by using these implemented databases, the development of informational tools, such as Air Quality Indices (AQIs), can be beneficial for presenting, in a comprehensible manner, new insight to policy makers and the public [22–24]. 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 aimed to present an ANN scheme for filling gaps in environmental and climate sciences and specifically in the field of air quality. ANNs usage for spatial-interpolation purposes is limited, and this work concentrates on the development of an effective method to spatially approximate air-quality parameters. From the original datasets and based on concentration time series for the selected pollutants of the EAQI, a shallow neural network 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. Data and Methodology

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, mild winters [25]. 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 [26–28]. The monitoring sites in the area are part of an air-quality-monitoring network that has operated since 1984, under the supervision of the Hellenic Ministry of Environment and Energy (MEE). Figure 1 presents the area of study and the locations of the monitoring sites (Table 1).

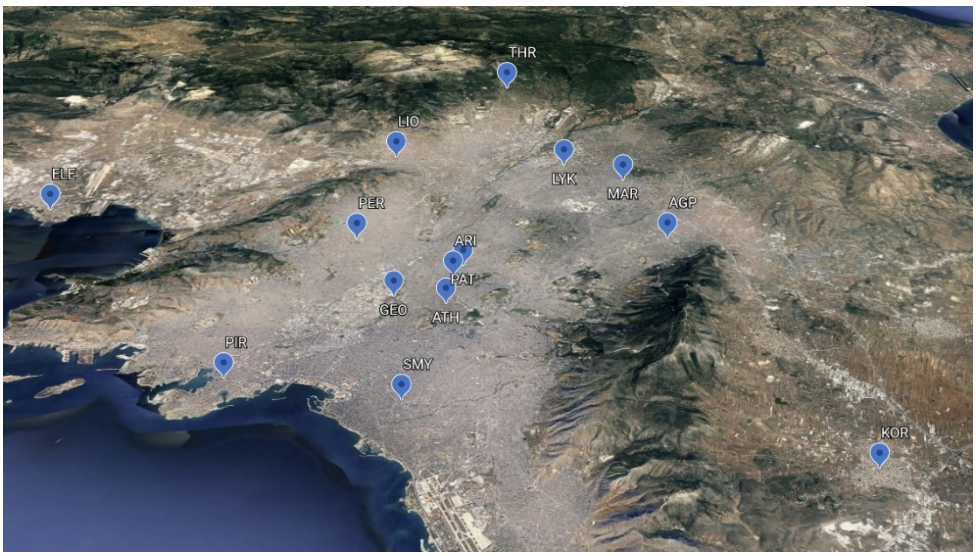

Figure 1. Spatial distribution of the air-quality-monitoring sites in the area of study.

Table 1. Air-quality-monitoring stations and their corresponding abbreviation and type.

Station	Abbreviation	Type
Ag. Paraskevi	AGP	Suburban/Background
Athinas	ATH	Urban/Traffic
Aristotelous	ARI	Urban/Traffic
Geoponiki	GEO	Suburban/Industrial
Elefsina	ELE	Suburban/Industrial
Thrakomakedones	THR	Suburban/Background
Koropi	KOR	Suburban/Background
Liosia	LIO	Suburban/Background
Lykovrisi	LYK	Suburban/Background
Marousi	MAR	Urban/Background
N. Smyrni	SMY	Urban/Background
Patission	PAT	Urban/Traffic
Piraeus	PIR	Urban/Traffic
Peristeri	PER	Urban/Background

The network is considered representative of the pollutants' spatial variability and, thus, suitable for the application of advanced statistical methodologies. As input data for the development of the neural network models, 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. The percentage of data availability for each station and pollutant was, in most cases, above 80%. However, the few exceptions for which the percentage was lower than 80% were also included in the analysis, as they could contribute to the interpolation process and, additionally, many of their missing concentrations could be targeted for data implementation. Only the stations that had no data availability for a whole year were excluded from this process. 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 also also a gap, it was not included in the interpolation process. This process was followed in order to avoid using a limited number of stations (or even an individual station) to interpolate missing values. The networks perform better when more information is provided. However, the same procedure could be performed by using less stations' data (and thus, not fulfilling the criterion that was mentioned before). In this case, less information would be available for the models in order to train, but more gaps could be filled, which would lead to a more complete database.

The results of the first step of the methodology are presented in Table 2 and reveal the number of missing values that can be potentially estimated initially and used to increase the available data points. The next step was to apply an 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 2).

Table 2. 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	2108	15.91
O_3	10,814	7961	2853	26.38
PM_{10}	7182	3948	3234	45.03
$PM_{2.5}$	4558	2524	2034	44.62
SO_2	7043	4746	2297	32.61

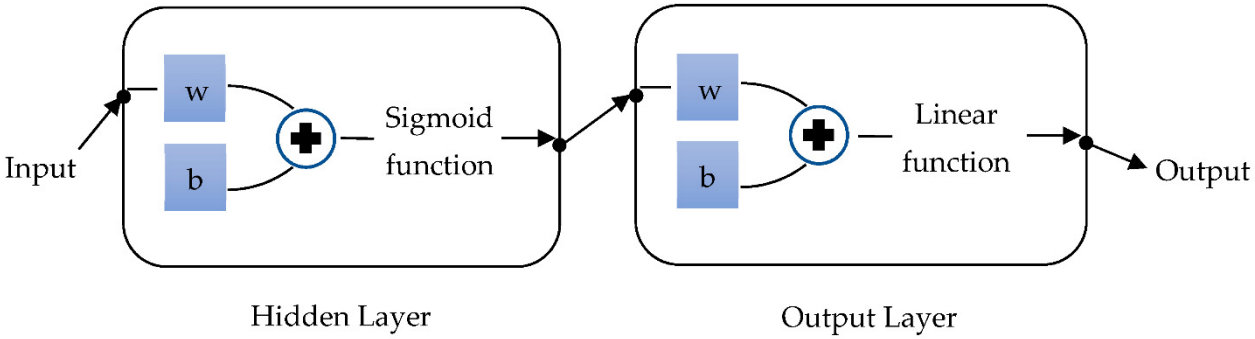

Figure 2. A two-layer network with sigmoid hidden neurons and linear output neurons.

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 and 15% testing). To reduce overfitting, the early stopping approach was utilized on the validation subset [26]. This approach terminates the training process when the validation subset's error begins to increase. Depending on the pollutant, the number of data points used for the subsets was different (as the number of stations with data availability is different) and is presented in Table 3. 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 [29–33], 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.

Table 3. 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	5412	5412	36,096
PM_{10}	13,410	2880	2880	19,170
$PM_{2.5}$	37,785	8100	8100	53,985
SO_2	13,925	3080	3080	20,085

Two more statistical metrics, the Root Mean Squared Error (RMSE) and the coefficient of determination (R^2), were also calculated, and in combination with MAE, they were used to provide a comparison between the results of the ANN methodology and a Multiple

Linear Regression (MLR) scheme. The MLR was applied according to the same criterion as with the ANN models. The equations for RMSE and R^2 are the following:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_i - O_i)^2} \quad (2)$$

$$R^2 = \left(\frac{\sum_{i=1}^{n}(O_i - \overline{O})(E_i - \overline{E})}{\sqrt{\sum_{i=1}^{n}(O_i - \overline{O})^2}\sqrt{\sum_{i=1}^{n}(E_i - \overline{E})^2}} \right)^2 \quad (3)$$

Lower MAE and RMSE values and higher R^2 illustrate the optimum performing scheme. Regarding the ANN method, five runs were performed for all models and for hidden layer neurons that ranged from 1 to 40. The best performing networks and their architecture are presented in Table 4. By using these selected SNN models for the corresponding inputs of 2018, the gaps in each station and pollutant were filled. Finally, on the interpolated datasets, mean and variance values were calculated and compared with the corresponding values of the original datasets for 2018.

Table 4. Number of input, hidden (average) and output neurons as well as the average Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and coefficient of determination (R^2) values and mean concentrations for the best performing models and the 2016/2017 time period. The MAE, RMSE and R^2 metrics include the results for the Multiple Linear Regression (MLR) scheme. For the MAE metric, the percentage of error (MAE to mean concentrations) was additionally calculated.

	Number of Neurons			MAE			RMSE			R^2	
	Input	Hidden	Output	Mean	ANNs	Error (%)	MLR	ANNs	MLR	ANNs	MLR
NO_2	13	21.7	1	32.70	5.80	17.74	7.23	8.31	9.87	0.76	0.67
O_3	12	22.3	1	58.86	6.86	11.65	9.32	9.78	12.49	0.87	0.77
PM_{10}	10	23.6	1	29.53	5.71	19.34	7.17	11.55	11.65	0.88	0.87
$PM_{2.5}$	5	25.2	1	23.81	5.17	21.71	5.68	8.47	8.97	0.69	0.65
SO_2	5	22.5	1	6.06	1.89	31.19	2.39	3.29	3.74	0.55	0.39

3. Results and Discussion

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 stations with data availability (inputs for the networks), 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 2 presents in detail the gaps originally and the number of them that will be eventually filled, after the interpolation, as well as the percentage of missing values that were estimated. It is noted that the number of gaps after the interpolation were calculated based on the criterion explained in the Methodology section and, thus, before the interpolation process, which provided the corresponding concentrations for each missing value.

The number of data points for the training, validation and testing subsets and for each pollutant is presented in Table 3. Pollutants with a 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 and 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.

The architecture of the optimum performance models is presented in Table 4. The hidden neurons number is an average of all the stations for each pollutant. The MAE, RMSE and R^2 average values (MAE and RMSE are 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. As mentioned beforehand, to illustrate the validity of the ANN approach, the steps of the analysis that were applied on the available datasets were also performed for MLR. Table 4 additionally presents the MAE, RMSE and R^2 results for the MLR method. It is evident that the ANNs are superior in all cases. The detailed results that include a station-by-station comparison are also provided in Supplementary Materials Tables S1–S5.

Tables 5–9 present the results for the mean and variance values of both the original and the gap-filled datasets for the five pollutants and the 2018 time period. It is noted that the differences are marginal in nearly all cases, and this is evident by the error value percentages (mean error and variance error).

Table 5. Mean and variance values results, for the original (*O*) and the interpolated (*I*) datasets, along with the corresponding error, per monitoring station for NO_2.

	Mean		**Mean Error**	**Variance**		**Variance Error**
	O	***I***		***O***	***I***	
AGP	14.06	14.06	-1.83×10^{-6}	147.91	147.86	-3.53×10^{-4}
ATH	44.24	42.90	-0.03	377.36	387.47	0.03
ARI	47.95	47.95	-5.35×10^{-5}	405.37	405.23	-3.45×10^{-4}
GEO	28.01	27.99	-7.81×10^{-4}	326.76	311.44	-0.05
ELE	24.41	24.42	2.99×10^{-4}	204.55	204.82	13×10^{-4}
THR	7.96	7.68	-0.04	94.85	87.34	-0.08
KOR	8.26	8.35	0.01	183.60	184.53	0.01
LIO	16.68	16.69	7.88×10^{-4}	187.13	187.86	39×10^{-4}
LYK	19.98	20.01	11×10^{-4}	286.24	284.08	-0.01
MAR	26.40	26.40	-8.63×10^{-6}	466.69	466.59	-2.17×10^{-4}
SMY	29.16	29.23	25×10^{-4}	480.56	483.84	0.01
PAT	70.95	70.94	-1.59×10^{-4}	750.28	750.33	6.07×10^{-5}
PIR	62.53	62.53	-4.43×10^{-5}	602.95	603.12	2.82×10^{-4}
PER	27.69	27.69	-7.61×10^{-5}	462.81	462.49	-6.80×10^{-4}

Table 6. Mean and variance values results, for the original (*O*) and the interpolated (*I*) datasets, along with the corresponding error, per monitoring station for O_3.

	Mean		**Mean Error**	**Variance**		**Variance Error**
	O	***I***		***O***	***I***	
AGP	82.61	82.66	5.05×10^{-4}	804.22	809.52	0.01
ATH	40.50	40.50	1.28×10^{-5}	882.74	882.54	-2.26×10^{-4}
GEO	56.27	59.01	0.05	1329.6	1334.4	36×10^{-4}
ELE	63.78	63.60	-29×10^{-4}	1226	1226	-6.03×10^{-5}
THR	96.50	96.51	1.24×10^{-4}	677.91	678.06	2.10×10^{-4}
KOR	65.77	65.77	-2.58×10^{-5}	607.77	608.73	16×10^{-4}
LIO	64.71	65.33	0.01	1199.8	1193.3	-0.01
LYK	64.42	64.07	-0.01	1352.1	1344.6	-0.01
MAR	65.81	65.88	11×10^{-4}	1387.2	1348.5	-0.03
SMY	73.65	73.87	30×10^{-4}	1314.2	1318.6	33×10^{-4}
PAT	16.52	16.83	0.02	278.29	296.92	0.07
PIR	40.45	40.43	-3.74×10^{-4}	944.15	943.88	-2.90×10^{-4}
PER	66.04	66.43	0.01	1385.3	1388.4	23×10^{-4}

By examining the total number of gaps in the original and interpolated databases of all pollutants (Table 2), it is evident that a considerable number of missing data points was estimated after the application of the methodology, which relies on the data-point availability of all the selected stations to interpolate the corresponding missing data points

(time-related) of the target station. In particular, the application of the ANNs, which utilized the available observations of the pollutants' concentrations based on the criterion that was introduced in the methodology, added a percent of missing values that ranged from about 16% to 45%, which depended on how many stations were used as inputs and the overall existing concentrations' distribution (whether for the same hour one or more stations had data availability). While the ANNs could be used to estimate data points at the target station, when not all stations had available data at the same hour, there are some factors that need to be considered. Although the ANNs provide representative estimations, there is always an associated error percentage when the results are compared with the observational data. This error percentage can be enhanced when the information provided at the models is not adequate for them to train effectively. If fewer stations were utilized, it could possibly lead to higher errors, and, in any case, the results should be analyzed carefully to find out how using a different number of inputs affects the output.

Table 7. Mean and variance values results, for the original (*O*) and the interpolated (*I*) datasets, along with the corresponding error, per monitoring station for PM_{10}.

	Mean		**Mean Error**	**Variance**		**Variance Error**
	O	***I***		***O***	***I***	
AGP	19.85	19.83	-7.09×10^{-4}	432.11	429.55	-0.01
ARI	36.38	36.37	-3.18×10^{-4}	630.70	628.58	-33×10^{-4}
ELE	29.27	29.06	-0.01	488.50	479.02	-0.02
THR	20.40	20.44	18×10^{-4}	414.02	401.35	-0.03
KOR	30.72	30.67	-16×10^{-4}	601.64	597.62	-0.01
LIO	33.74	32.93	-0.02	806.56	704.83	-0.13
LYK	27.03	27.63	0.02	378.55	657.21	-0.74
MAR	29.48	29.48	7.46×10^{-5}	685.14	680.14	-0.01
SMY	31.03	30.79	-0.01	666.01	644.88	-0.03
PIR	39.34	39.49	37×10^{-4}	695.72	697.59	27×10^{-4}
PER	30.31	30.32	2.16×10^{-4}	713.39	707.80	-0.01

Table 8. Mean and variance values results, for the original (*O*) and the interpolated (*I*) datasets, along with the corresponding error, per monitoring station for $PM_{2.5}$.

	Mean		**Mean Error**	**Variance**		**Variance Error**
	O	***I***		***O***	***I***	
AGP	11.60	11.60	-4.47×10^{-4}	42.35	42.17	-42×10^{-4}
ARI	19.11	18.92	-0.01	213.67	204.40	-43×10^{-4}
ELE	17.81	17.83	9.18×10^{-4}	100.90	99.98	-92×10^{-4}
THR	13.44	13.43	-10×10^{-4}	45.47	44.63	-186×10^{-4}
LYK	15.28	15.47	0.01	133.63	133.81	14×10^{-4}
PIR	18.00	18.24	0.01	178.46	183.70	29×10^{-4}

Table 9. Mean and variance values results, for the original (*O*) and the interpolated (*I*) datasets, along with the corresponding error, per monitoring station for SO_2.

	Mean		**Mean Error**	**Variance**		**Variance Error**
	O	***I***		***O***	***I***	
ATH	4.21	4.21	15×10^{-4}	12.22	12.29	53×10^{-4}
ARI	4.46	4.58	265×10^{-4}	11.57	11.70	109×10^{-4}
ELE	10.73	10.57	-150×10^{-4}	45.35	47.76	529×10^{-4}
KOR	4.92	4.91	-25×10^{-4}	7.49	7.45	-51×10^{-4}
PAT	8.87	8.85	-28×10^{-4}	20.17	20.20	14×10^{-4}
PIR	9.93	9.92	-13×10^{-4}	66.21	65.52	-105×10^{-4}

As proposed by Willmott and Matsuura [29], dimensioned evaluations of model-performance error should be based on MAE. However, a better understanding of the MAE values can be achieved by calculating the percentage of error (MAE to mean concentration). According to Table 4 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 3 ($PM_{2.5}$ has nearly three times more data points). Different approaches to evaluate the performance of the models can be followed (scatter diagrams, etc.), and more types of similar complexity neural network models can be examined.

4. 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 indices 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/climatic parameters as inputs.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/atmos12040499/s1, Table S1:NO_2 performance statistics (MAE, RMSE and R^2 results) for the FFNNs and MLR schemes, Table S2: O_3 performance statistics (MAE, RMSE and R^2 results) for the FFNNs and MLR schemes, Table S3: PM_{10} performance statistics (MAE, RMSE and R^2 results) for the FFNNs and MLR schemes, Table S4: $PM_{2.5}$ performance statistics (MAE, RMSE and R^2 results) for the FFNNs and MLR schemes, Table S5: SO_2 performance statistics (MAE, RMSE and R^2 results) for the FFNNs and MLR schemes.

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; individually, C.G.T. was responsible for the data curation and validation of the results and supervised the whole procedure. All authors (C.G.T., A.A. and I.K.) 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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated during and/or analyzed during the current study are publicly available in the Ministry of Environment and Energy repository (ypen.gov.gr).

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

References

1. Amanollahi, J.; Tzanis, C.; Abdullah, A.M.; Ramli, M.F.; Pirasteh, S. Development of the models to estimate particulate matter from thermal infrared band of Landsat Enhanced Thematic Mapper. *Int. J. Environ. Sci. Technol.* **2013**, *10*, 1245–1254. [CrossRef]
2. Baklanov, A.; Molina, L.T.; Gauss, M. Megacities, air quality and climate. *Atmos. Environ.* **2016**, *126*, 235–249. [CrossRef]
3. European Environment Agency. *Air Quality in Europe—2013 Report: EEA Report No. 9/2013*; European Union: Luxembourg, 2013. Available online: http://www.eea.europa.eu/publications/air-quality-in-europe-2013 (accessed on 11 November 2020).
4. Grøntoft, T. Estimation of damage cost to building façades per kilo emission of air pollution in Norway. *Atmosphere* **2020**, *11*, 686. [CrossRef]
5. de la Fuente, D.; Vega, J.M.; Viejo, F.; Díaz, I.; Morcillo, M. City scale assessment model for air pollution effects on the cultural heritage. *Atmos. Environ.* **2011**, *45*, 1242–1250. [CrossRef]
6. Can, A.; Dekoninck, L.; Botteldooren, D. Measurement network for urban noise assessment: Comparison of mobile measurements and spatial interpolation approaches. *Appl. Acoust.* **2014**, *83*, 32–39. [CrossRef]
7. Denby, B.; Sundvor, I.; Cassiani, M.; de Smet, P.; de Leeuw, F.; Horálek, J. Spatial Mapping of Ozone and SO2 Trends in Europe. *Sci. Total Environ.* **2010**, *408*, 4795–4806. [CrossRef] [PubMed]
8. Li, J.; Heap, A.D. Spatial interpolation methods applied in the environmental sciences: A review. *Environ. Model. Softw.* **2014**, *53*, 173–189. [CrossRef]
9. Liang, F.; Xiao, Q.; Wang, Y.; Lyapustin, A.; Li, G.; Gu, D.; Pan, X.; Liu, Y. MAIAC-based long-term spatiotemporal trends of PM2.5 in Beijing, China. *Sci. Total Environ.* **2018**, *616–617*, 1589–1598. [CrossRef]
10. Yang, J.; Hu, M. Filling the missing data gaps of daily MODIS AOD using spatiotemporal interpolation. *Sci. Total Environ.* **2018**, *633*, 677–683. [CrossRef]
11. Zhang, R.; Di, B.; Luo, Y.; Deng, X.; Grieneisen, M.L.; Wang, Z.; Yao, G.; Zhan, Y. A nonparametric approach to filling gaps in satellite-retrieved aerosol optical depth for estimating ambient PM2.5 levels. *Environ. Pollut.* **2018**, *243*, 998–1007. [CrossRef] [PubMed]
12. Blanchard, C.L.; Tanenbaum, S.; Hidy, G.M. Spatial and temporal variability of air pollution in Birmingham, Alabama. *Atmos. Environ.* **2014**, *89*, 382–391. [CrossRef]
13. Ma, J.; Cheng, J.C.P.; Lin, C.; Tan, Y.; Zhang, J. Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. *Atmos. Environ.* **2019**, *214*, 116885. [CrossRef]
14. Qi, Y.; Li, Q.; Karimian, H.; Liu, D. A hybrid model for spatiotemporal forecasting of PM 2.5 based on graph convolutional neural network and long short-term memory. *Sci. Total Environ.* **2019**, *664*, 1–10. [CrossRef] [PubMed]
15. Vakili, M.; Sabbagh-Yazdi, S.R.; Khosrojerdi, S.; Kalhor, K. Evaluating the effect of particulate matter pollution on estimation of daily global solar radiation using artificial neural network modeling based on meteorological data. *J. Clean. Prod.* **2017**, *141*, 1275–1285. [CrossRef]
16. Zainuddin, Z.; Pauline, O. Modified wavelet neural network in function approximation and its application in prediction of time-series pollution data. *Appl. Soft Comput. J.* **2011**, *11*, 4866–4874. [CrossRef]
17. Coman, A.; Ionescu, A.; Candau, Y. Hourly ozone prediction for a 24-h horizon using neural networks. *Environ. Model. Softw.* **2008**, *23*, 1407–1421. [CrossRef]
18. Chattopadhyay, S. Feed forward Artificial Neural Network model to predict the average summer-monsoon rainfall in India. *Acta Geophys.* **2007**, *55*, 369–382. [CrossRef]
19. Wahid, H.; Ha, Q.P.; Duc, H.; Azzi, M. Neural network-based meta-modelling approach for estimating spatial distribution of air pollutant levels. *Appl. Soft Comput. J.* **2013**, *13*, 4087–4096. [CrossRef]
20. Cheng, J.C.P.; Ma, L.J. A data-driven study of important climate factors on the achievement of LEED-EB credits. *Build. Environ.* **2015**, *90*, 232–244. [CrossRef]
21. Yang, Z.; Wang, J. A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction. *Environ. Res.* **2017**, *158*, 105–117. [CrossRef] [PubMed]
22. Zhan, D.; Kwan, M.P.; Zhang, W.; Yu, X.; Meng, B.; Liu, Q. The driving factors of air quality index in China. *J. Clean. Prod.* **2018**, *197*, 1342–1351. [CrossRef]
23. Silva, L.T.; Mendes, J.F.G. City Noise-Air: An environmental quality index for cities. *Sustain. Cities Soc.* **2012**, *4*, 1–11. [CrossRef]
24. Ganguly, N.D.; Tzanis, C.G.; Philippopoulos, K.; Deligiorgi, D. Analysis of a severe air pollution episode in India during Diwali festival—A nationwide approach. *Atmósfera* **2019**, *32*, 225–236. [CrossRef]
25. Tzanis, C.G.; Koutsogiannis, I.; Philippopoulos, K.; Deligiorgi, D. Recent climate trends over Greece. *Atmos. Res.* **2019**, *230*, 104623. [CrossRef]
26. Tzanis, C.G.; Alimissis, A.; Philippopoulos, K.; Deligiorgi, D. Applying linear and nonlinear models for the estimation of particulate matter variability. *Environ. Pollut.* **2019**, *246*, 89–98. [CrossRef]
27. Varotsos, C.; Christodoulakis, J.; Tzanis, C.; Cracknell, A.P. Signature of tropospheric ozone and nitrogen dioxide from space: A case study for Athens, Greece. *Atmos. Environ.* **2014**, *89*, 721–730. [CrossRef]
28. Tzanis, C.; Varotsos, C.A. Tropospheric aerosol forcing of climate: A case study for the greater area of Greece. *Int. J. Remote Sens.* **2008**, *29*, 2507–2517. [CrossRef]
29. Willmott, K.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82. [CrossRef]

30. Alimissis, A.; Philippopoulos, K.; Tzanis, C.G.; Deligiorgi, D. Spatial estimation of urban air pollution with the use of artificial neural network models. *Atmos. Environ.* **2018**, *191*, 205–213. [CrossRef]
31. Fallahi, S.; Amanollahi, J.; Tzanis, C.G.; Ramli, M.F. Estimating solar radiation using NOAA/AVHRR and ground measurement data. *Atmos. Res.* **2018**, *199*, 93–102. [CrossRef]
32. Rahimpour, A.; Amanollahi, J.; Tzanis, C.G. Air quality data series estimation based on machine learning approaches for urban environments. *Air Qual. Atmos. Health* **2021**, *14*, 191–201. [CrossRef]
33. Mirzaei, M.; Amanollahi, J.; Tzanis, C.G. Evaluation of linear, nonlinear, and hybrid models for predicting $PM_{2.5}$ based on a GTWR model and MODIS AOD data. *Air Qual. Atmos. Health* **2019**, *12*, 1215–1224. [CrossRef]