Hindawi
Complexity
Volume 2018, Article ID 7238015, 14 pages
https://doi.org/10.1155/2018/7238015

*Research Article*

# Neural Models for Imputation of Missing Ozone Data in Air-Quality Datasets

**Ángel Arroyo,[1] Álvaro Herrero,[1] Verónica Tricio,[2] Emilio Corchado,[3] and Michał Woźniak[4]**

[1] *Department of Civil Engineering, University of Burgos, Burgos, Spain*
[2] *Department of Physics, University of Burgos, Burgos, Spain*
[3] *Departamento de Informática y Automática, University of Salamanca, Salamanca, Spain*
[4] *Department of Systems and Computer Networks, Wrocław University of Science and Technology, Wrocław, Poland*

Correspondence should be addressed to Ángel Arroyo; aarroyop@ubu.es

Received 5 December 2017; Accepted 31 January 2018; Published 8 March 2018

Academic Editor: Eloy Irigoyen

Ozone is one of the pollutants with most negative effects on human health and in general on the biosphere. Many data-acquisition networks collect data about ozone values in both urban and background areas. Usually, these data are incomplete or corrupt and the imputation of the missing values is a priority in order to obtain complete datasets, solving the uncertainty and vagueness of existing problems to manage complexity. In the present paper, multiple-regression techniques and Artificial Neural Network models are applied to approximate the absent ozone values from five explanatory variables containing air-quality information. To compare the different imputation methods, real-life data from six data-acquisition stations from the region of Castilla y León (Spain) are gathered in different ways and then analyzed. The results obtained in the estimation of the missing values by applying these techniques and models are compared, analyzing the possible causes of the given response.

## 1. Introduction and Related Work

The ozone ($O_3$) is an odorless, colorless, and highly reactive gas composed of three oxygen atoms. It is formed both in the Earth's upper atmosphere (stratospheric ozone) and at ground level (tropospheric ozone). It can be "good" or "bad" for people's health and for the environment, depending on its concentration levels and location in the atmosphere [1].

Stratospheric $O_3$ is formed naturally through the interaction of solar UltraViolet (UV) radiation with molecular oxygen ($O_2$). Ground-level or "bad" ozone is not emitted directly into the air. In the 1950s, hydrocarbons and nitrogen oxides ($NO_x$) were identified as the two key chemical precursors of photochemical smog and its concomitant high concentrations of $O_3$ and other photochemical oxidant [2]. The majority of ground-level $O_3$ is formed from the photochemical oxidation of Volatile Organic Compounds (VOCs) in the presence of NO and other $NO_x$. Significant sources of VOCs are chemical plants, gasoline pumps, oil-based paints, autobody shops, and print shops. $NO_x$ result primarily from high temperature combustion, and its most significant sources are power plants, industrial furnaces and boilers, and motor vehicles [3].

*1.1. Importance of Ozone.* The $O_3$ exposition can cause damage in different ways. In the stratosphere, reduced $O_3$ levels as a result of $O_3$ layer depletion mean less protection from the sun's rays and more exposure to UltraViolet B (shortwave) rays (UVB) radiation at the Earth's surface [4]. The effects on human health of the $O_3$ layer depletion have been much analyzed, increasing the amount of UVB that reaches the Earth's surface. UVB causes nonmelanoma skin cancer and plays a major role in malignant melanoma development. In addition, UVB has been linked to the development of certain cataracts, negative effects in patients with asthma, and other chronic respiratory disease. With respect to ground-level $O_3$, and its effects on human health, breathing $O_3$ can trigger a variety of health problems. People with asthma and other chronic respiratory disease are a large and growing segment of the population and are also known to be especially susceptible to the effects of $O_3$ exposure. On days with high levels of $O_3$, people with asthma tend to experience increased respiratory symptoms [3]. The layer $O_3$ depletion has also negative effects on the process of the development of plants, effects on the marine ecosystems like a direct reduction in phytoplankton production, negative effects on materials like biopolymers, and so forth. Tropospheric $O_3$ does not provide the protective function that it fulfills in the stratosphere, being high reactivity. Its strong oxidizing capacity, when its levels rise above the natural background, can cause adverse effects in materials (derived from its corrosive effects), on vegetation and ecosystems.

The present work focuses on tropospheric $O_3$, which is a risk for the air quality [3]. Given the increase in $O_3$ levels in the troposphere, it is currently considered one of the most important atmospheric pollutants.

### 1.2. Ozone Level Monitoring.

Around the world there are numerous data-acquisition networks for the measurement of $O_3$ levels and other pollutants, which consist of many stations in different locations where different sensors measure corresponding magnitudes. These network stations acquire data at periodic intervals of time (periods between ten and fifteen minutes are the most frequent ones) but frequently appear missing or corrupted data. In Europe, data are considered as corrupted when not meeting the Council Decision 97/101/EC of January 27, 1997 [5], which establish a reciprocal exchange of information and data from networks and individual stations measuring ambient air pollution within the Member States. Some of these networks provide information about the validity of the data, indicating through codes if the data is correct, it has not been possible to acquire, or it is corrupt, but in other occasions this type of information is not provided while the data are still missing. Some reasons for such failures have been pinpointed [6], namely, a damaged cable, the loss of proper electrical grounding, half-melted frost or snow on the dome, communications failure, and so forth. Some of these causes are temporary and may disappear spontaneously, but other ones require the intervention of a maintenance task force, and therefore errors persist for different periods of time. The absence of valid data may also be due to reasons such as the following: mishandling of samples, low signal-to-noise ratio, measurement error, nonresponse, or deleted aberrant value [7]. This is a problem for the analysis of the information coming from the measurement networks, and the imputation of these missing data [8] is necessary. Any of the variables acquired in network stations may suffer from the problem of the absence of data. If many data variables are omitted or corrupted in the same record, the whole sample must be withdrawn, when some models are applied [9], for subsequent tasks such as control, classification, forecast. Alternatively, if data for the same pollutant are missing in several adjacent rows, removing that variable may also be an alternative solution. In conclusion, having a complete set of data is necessary to perform a reliable study and to apply some models that cannot deal with missing data.

### 1.3. Missing Values and Related Work.

The standard classification of missing data phenomenon [10] includes different situations:

(i) Missing Completely At Random (MCAR), when the probability of an instance (case) having a missing value for a variable does not depend on either the known values or the missing data.

(ii) Missing At Random (MAR), when the probability of an instance having a missing value for a variable may depend on the known values but not on the value of the missing data itself.

(iii) Not Missing At Random (NMAR), when the probability of an instance having a missing value for a variable could depend on the value of that variable.

As previous authors have pointed out, the complexity varies between these patterns of missing data [11]. Usually, in the case of air-quality data, missing values are associated with MAR or MCAR. The circumstances that may interfere with the acquisition of the data are many and not easily predictable [12].

To solve the missing data problem, a wide variety of different methods have been applied up to now [8, 10, 13]. These imputation methods (IMs) are usually classified as follows:

(i) Single imputation (SI): the method fills in one value for each missing one [12].

(ii) Multiple imputation (MI): multiple simulated values are generated at the same time [14].

The univariate and multivariate imputation methods differ in which the approximation of the missing values of the variable under study are calculated from the rest of the values of the very same variable (univariate) or using values of the rest of the variables (multivariate) [12].

With the aim of reducing the complexity of other MI applied methods [11], the present paper focuses on single and multivariate imputation for the $O_3$ magnitude in air pollution datasets. To do so, multiple-regression (linear and nonlinear) techniques together with Artificial Neural Networks (ANN) are applied to real-life datasets obtained from public air-quality networks.

Up to now, different Artificial-Intelligence (AI) techniques have been applied for imputation of missing data. In [7] imputation methods based on six different techniques are compared: $K$-Nearest Neighbors (KNN), Fuzzy $K$-Means (FKM), Singular Value Decomposition, Bayesian Principal Component Analysis (bPCA) and Multiple Imputations by Chained Equations. These methods are applied to four datasets split into two groups of various sizes: small datasets (Iris and *E. coli*) and large datasets (breast cancers 1 and 2). bPCA and FKM appeared to be the most robust imputation methods in the tested conditions.

In [15] the accuracy of different imputation methods is evaluated: MissForest (MF) and Multiple Imputation based on Expectation-Maximization (MIEM), along with two other imputation methods: Sequential Hot-Deck and Multiple Imputation based on Logistic Regression (MILR). The models are applied over fourteen binary datasets, with a range of missing data rates between 5% and 50%. The results from 10-fold Cross-Validation (CV) show that the performance of the imputation methods varies substantially between different classifiers and at different rates of missing values.

Although many imputation methods have been proposed up to now, scant attention has been paid to validate ANN for such a task, taking advantage of their regression capability [16]. Among these previous studies, ANN have been applied for the estimation of lost values in [17], where the main goal is identifying Learning Disabilities (LD) in children at early stages. In [18], authors proposed a SI approach relying on a Multilayer Perceptron (MLP) whose training is conducted with different learning rules, and a MI approach based on the combination of MLP and KNN. 24 real and simulated datasets from the UCI repository, the Promise repository, and mldata.org were exposed to a perturbation experiment with random generation of monotone missing data pattern.

In [19] six different types of ANN are proposed as IM: MLP and its variations (the Time-Lagged Feedforward Network (TLFN)), the Generalized Radial-Basis-Function (GRBF) network, the Recurrent Neural Network (RNN), and its variations (the Time Delay Recurrent Neural Network (TDRNN)). Additionally, the Counterpropagation Fuzzy-Neural Network (CFNN) along with different optimization methods is applied for infilling missing daily total precipitation and extreme temperature series from 15 weather stations. The standard MLP and TLFN appear to provide the most accurate reconstruction of missing precipitation and daily extreme temperatures records with results for the $R$ correlation coefficient between the observed and the reconstructed daily series close to 1.

In [20] a novel nonparametric algorithm named Generalized regression neural network Ensemble for Multiple Imputation (GEMI) is proposed. Additionally, a SI version of this approach (GESI) is proposed. The algorithms were tested on 98 synthetic and real-world datasets. All simulation results show the advantages of GEMI as compared with conventional algorithms. GEMI has heavy memory storage requirements but outperformed other SI algorithms.

In [21] fifteen real and simulated datasets are exposed to a perturbation experiment, based on the random generation of missing values. Several architectures and learning algorithms for the MLP are tested and compared with three classic imputation procedures: mean/mode imputation, regression, and hot-deck [22].

In [23] a methodology based on Gaussian Mixture Model (GMM) and Extreme Learning Machine (ELM) is developed and tested on some datasets from the UCI Machine Learning Repository and the LIACC regression repository. GMM is used to model the data distribution which is adapted to handle missing values, while ELM enables devising a Multiple Imputation strategy for final estimation. The combination of GMM and ELM is shown to be superior in almost all tested cases over the method based on conditional mean imputation.

In [24] a SI approach relying on a MLP and a MI approach based on the combination of MLP and $K$-NN is proposed. The models are applied to 18 real and simulated datasets like domains such as biology, medicine, chemistry, electronics, social surveys, census, and business. For datasets with only quantitative variables MIMLP model provided the best results, with IMLP being the best method for datasets with categorical variables.

In [25] a two-stage hybrid model for filling the missing values using fuzzy c-means clustering and MLP is proposed. It is applied to a Wine dataset with a 1% to 5% of generated missing values and the accuracy of the model is checked using the Mean Absolute Percentage Error (MAPE). The MAPE obtained for stage 2 (MLP regression to the obtained dataset as a result of applying fuzzy $c$-means in stage 1) is 4.95% for 1% missing-value records and 8.36% for 5% missing-value records.

In the case of air-quality data, few imputation methods have been proposed up to now. In [13], an important set of SI: Listwise, Unconditional mean, Modified Median, Principal Component-based, Expectation-Maximization (EM) (Regularized-EM), and MI methods are applied to three datasets with the most important pollutant variables (NO, $NO_2$, $NO_x$, CO, $O_3$, PM10, and PM2.5) and a percentage of missing data among the 3.85% and the 23.52% depending on the year. Missing data of the eight variables are imputed in order to assess the effectiveness of the methods applied. In general, MI tends to yield more scattered values than its counterparts, mainly when the variables have many voids and they correlate poorly to the other variables like CO with 43.5% of missing data in 2006 and they correlate poorly to the other variables.

In [11] some methods for the imputation of missing air-quality data are compared: in the context of SI (linear, spline, and nearest neighbor interpolations), MI (regression-based imputation, multivariate nearest neighbor, Self-Organizing Maps (SOM), and Multilayer Backpropagation (MLBP) nets) and hybrid methods of the aforementioned. The dataset uses the most common pollutants: $NO_x$, $NO_2$, $O_3$, PM10, $SO_2$, and CO concentrations, all on a time-scale of one per hour (hourly averaged), together with four meteorological parameters. The performance of the proposed univariate missing data interpolation was limited, and in general they were able to fill only very short gaps of contiguous missing data. The general performance of the applied imputation methods was fair good when considering the pollutants ($NO_x$, $NO_2$, $O_3$, PM10, $SO_2$, and CO) which are the most important ones in terms of air-quality modelling, but not so good regarding meteorological variables. The results suggested that SOM and MLBP are the methods of choice for air-quality data imputation and even better results can be achieved by using the MI.

*1.4. Main Contributions.* The main contributions of this work are as follows:

(i) Deep study of the real-life human health protection task in Spanish region of Castilla y León.

(ii) Multisensor of $O_3$ data analysis.

(iii) Experimental evaluation of the proposed approach based on multiple-regression techniques together with ANN models.

To the best of authors knowledge, this is the first approach of imputation methods of $O_3$ based on both MLP and Radial-Basis-Function Networks.

The rest of this paper is organized as follows. Section 2 presents the techniques and models applied. Section 3 details the real-life case study that is addressed in present work, while Section 4 describes the experiments and results. Finally, Section 5 sets out the main conclusions and future work.

## 2. Regression Techniques and ANN Models

In order to fill missing or corrupted values of $O_3$ in high dimensional datasets with air-quality information, two regression techniques and two ANN models have been applied in present study. This set of techniques applied as imputation methods is described in this section.

*2.1. Regression Techniques.* Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable [26].

The general purpose of multiple regressions [27] is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable.

*2.1.1. Multiple Linear Regression.* Multiple linear regression (MLR) attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data [28]. Every value of the independent variable ($x$) is associated with a value of the dependent variable ($y$). The population regression line for $p$ explanatory variables

$$x_1, x_2, \ldots, x_p \tag{1}$$

is defined to be

$$u_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \tag{2}$$

This line describes how the mean response $u_y$ changes with the explanatory variables. The observed values for $y$ vary about their means $u_y$ and are assumed to have the same standard deviation $\sigma$. The fitted values $b_0, b_1, \ldots, b_p$ estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ of the population regression line.

Since the observed values for $y$ vary about their means $u_y$, the multiple-regression models include a term for this variation. The model is expressed as DATA = FIT + RESIDUAL, where the "FIT" term represents the expression $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. The "RESIDUAL" term represents the deviations of the observed values $y$ from their means $u_y$, which are normally distributed with mean 0 and variance $\sigma$. The notation for the model deviations is $\varepsilon$.

Formally, the model for multiple linear regression, given $n$ observations, is [28]

$$Yi = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad \text{for } i = 1, 2, \ldots, n. \tag{3}$$

*2.1.2. Multiple Nonlinear Regression.* A Multiple Nonlinear Regression (MN-LR) is a form of regression analysis in which observational data are modelled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables [29]. The data are fitted by a method of successive approximations.

The parameters can take the form of an exponential, trigonometric, power, or any other nonlinear function. To determine the nonlinear parameter estimates, an iterative algorithm is typically used.

$$y = f(X, B) + \varepsilon, \tag{4}$$

where $B$ represents nonlinear parameter estimates to be computed, $X$ is the dependent or criterion variables, and $\varepsilon$represents the error terms.

*2.2. Artificial Neural Networks.* Artificial Neural Networks (ANN), also known as Artificial Neural Systems (ANS), connectionist systems, adaptive networks, and distributed and parallel processing are simplified models of natural neural systems. The following definition, given by Hecht-Nielsen in 1989 [30], formalizes the concept of ANN:

> *An ANN is a parallel processing computer system distributed, consisting of a set of elementary processing units equipped with a small local memory and interconnected in a network through connections with associated weights. Each processing unit has one or more input connections and a single output connection that links to many collateral connections as desired. All processing associated with an elementary unit is a local, i.e. depends only on the values that take input signals from the unit and the internal state of the same.*

*2.2.1. Multilayer Perceptron (MLP).* The MLP consists of a system of simple interconnected neurons or nodes. The nodes are connected by weights and output signals which are a function of the sum of the inputs to the node modified by a simple nonlinear transfer, or activation, function. The architecture consists of several layers of neurons; the input layer serves to pass the input vector to the network. The terms "input vectors" and "output vectors" refer to the inputs and outputs of the MLP and can be represented as single vectors [31]. A MLP may have one or more hidden layers and finally an output layer. MLP are fully connected, with each node connected to every node in the next and previous layer.

To perform a comprehensive comparison, the MLP is trained with the following algorithms:

(1) Levenberg-Marquardt backpropagation (LM)

(2) Gradient Descent with momentum and adaptive learning rate backpropagation (GDX) [32]

(3) Batch Training with weight and bias learning rules (TB)

(4) Scaled Conjugate Gradient backpropagation (SCG)

(5) Bayesian Regularization backpropagation (BR).

*2.2.2. Radial-Basis-Function Networks (RBFN).* In a RBFN [33] each unit in the hidden layer of this network has its own centroid, and, for each input vector $x = (x_l, x_2, \ldots, x_n)$, it computes the distance between $x$ and its centroid. Its output of the unit is calculated as a nonlinear function of this distance.

Assuming that there are $r$ input nodes and $m$ output nodes, the overall response function without considering nonlinearity in an output node has the following form [34]:

$$\sum_{i=1}^{M} W_i * K\left(\frac{x - z_i}{\sigma_i}\right) = \sum_{i=1}^{M} W_i * g\left(\frac{\left\|x - z_i\right\|}{\sigma_i}\right), \qquad (5)$$

where $M \in \mathbb{N}$ is the number of units in the hidden layer, $W_i \in \mathbb{R}^m$ is the vector of weights linking the $i$th hidden-layer unit to the output nodes, $x$ is an input vector, $K$ is a radially symmetric kernel function of a unit in the hidden layer, $z_i$ and $\sigma_i$ are the centroid and smoothing factor of the $i$th kernel node, respectively, and $g$: $[0, \infty) \rightarrow \mathbb{R}$ is a function called the activation function, which characterizes the kernel shape.

## 3. Case Study

In present study, data from air-quality stations in Castilla y León (CyL) are analyzed. CyL is a Spanish region located at the north-center of the Iberian Peninsula. It is composed of nine provinces and it is the most extensive region of Spain with a total surface of 94,226 square kilometers and the sixth with more population: 2,435,797 habitants. Gross Domestic Product (GDP) in CyL represents the 5.3% of country's GDP [35]. Climate in CyL approaches what is known as the continental ocean, characterized by cold winters and hot summers with short spring and autumn periods.

CyL region provides a wide network of stations [36] for the acquisition of air-quality data. These data are public available according to the Open Data Initiative from the Spanish Government [37].

Stations from this network have some interesting characteristics:

(1) Stations are classified in types: urban, background, and oriented to the vegetation protection [36].

(2) These stations collect the fundamental air-quality pollutants, and among them is the $O_3$, which is the objective pollutant of this study. Daily averages data [38] of each pollutant are provided in each location.

(3) This data presents empty or corrupted data in all of its variables in some rows and in a reasonable percentage to be estimated.

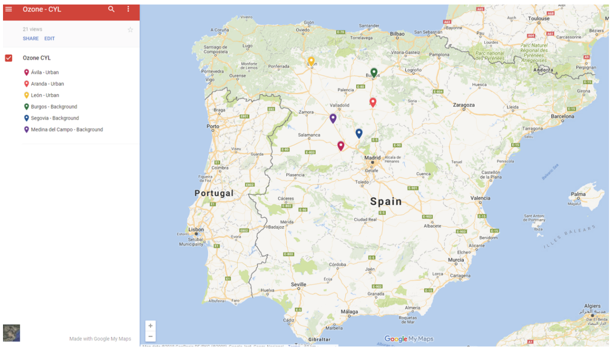

Figure 1: Location of the six selected stations in CyL, by Google Maps.

In the present study, pollutant data recorded in six different stations from the CyL network are analyzed. Daily data averages from years 2000 to 2008 have been selected. For some periods of time within the selected time window, data are not available for all the variables and, thus, the whole example is rejected for the study. Three of the stations are located in the center of the cities and labeled as urban stations; these stations are oriented to the protection of the human health. The other three stations are background stations and are also oriented to the protection of the human health. These stations measure a greater number of pollutants than the other type of stations and are the most important ones in terms of air quality, and many of them are not collected at the stations for the vegetation protection. This fact is important for the determination of the $O_3$ missing values, as this gas is especially harmful for human health.

The three urban stations considered in present study are as follows:

(1) Ávila. "Bus Station" station. Geographical coordinates: 40.65914, −4.68237; 1150 meters above sea level (masl).

(2) Aranda de Duero. "Jardines de Don Diego" station. Geographical coordinates: 41.67111, −3.68388; 801 masl.

(3) León. "Avda. San Ignacio de Loyola" station. Geographical coordinates: 42.60388, −5.58722; 838 masl.

The three background stations are as follows:

(1) Burgos. "Fuentes Blancas" station. Geographical coordinates: 42.33611, −3.63611; 929 masl.

(2) Segovia. "Acueducto" station. Geographical coordinates: 40.95555, −4.11055; 951 masl.

(3) Medina del Campo (Valladolid). "Bus Station" station. Geographical coordinates: 41.31638, −4.90916; 721 masl.

Figure 1 shows the location of the six selected stations that have been studied in the present paper.

The pollutants gathered in the above-mentioned stations and analyzed in the present study are as follows:

Table 1: Correlation matrix of the six variables in the dataset.

| | $O_3$ | CO | NO | $NO_2$ | PM10 | $SO_2$ |
|---|---|---|---|---|---|---|
| **$O_3$** | **1.000** | **−0.123** | **−0.161** | **−0.202** | **0.072** | **−0.013** |
| CO | −0.123 | 1.000 | 0.360 | 0.412 | 0.358 | 0.299 |
| NO | −0.161 | 0.360 | 1.000 | 0.540 | 0.233 | 0.330 |
| $NO_2$ | −0.202 | 0.412 | 0.540 | 1.000 | 0.330 | 0.257 |
| PM10 | 0.072 | 0.358 | 0.233 | 0.330 | 1.000 | 0.251 |
| $SO_2$ | −0.013 | 0.299 | 0.330 | 0.257 | 0.251 | 1.000 |

Table 2: Percentage of missing and corrupted data for each one of the analyzed variables.

| | $O_3$ | NO | $NO_2$ | CO | PM10 | $SO_2$ |
|---|---|---|---|---|---|---|
| Missing | **8.104%** | 8.020% | 8.034% | 8.554% | 9.131% | 8.196% |
| Corrupted | **1.857%** | 2.047% | 1.815% | 2.926% | 2.413% | 1.801% |
| *Total* | **9.961%** | 10.067% | 9.849% | 11.480% | 11.544% | 9.997% |

(1) Ozone ($O_3$), $\mu$g/m$^3$, secondary pollutant. See Section 1.

(2) Carbon monoxide (CO), mg/m$^3$, primary pollutant. It is an odorless, colorless gas formed by the incomplete combustion of fuels. When people are exposed to CO gas, the CO molecules will displace the oxygen in their bodies and lead to poisoning [39].

(3) Nitric oxide (NO), $\mu$g/m$^3$, primary pollutant. NO is a colorless gas which reacts with ozone undergoing rapid oxidation to $NO_2$, predominant in the atmosphere [39].

(4) Nitrogen dioxide ($NO_2$), $\mu$g/m$^3$, primary pollutant. From the standpoint of health protection, nitrogen dioxide has set exposure limits for long and short duration [39].

(5) Particulate matter (PM10), $\mu$g/m$^3$, primary pollutant. These particles remain stable in the air for long periods of time without falling to the ground and can be moved significant distances by the wind. It is defined by the ISO as follows: "particles which pass through a size-selective inlet with a 50% efficiency cut-off at 10 $\mu$m aerodynamic diameter. PM10 corresponds to the 'thoracic convention' as defined in ISO 7708:1995, Clause 6" [40].

(6) Sulphur dioxide ($SO_2$), $\mu$g/m$^3$, primary pollutant. It is a gas. It smells like burnt matches. Its smell is also suffocating. $SO_2$ is produced by volcanoes and in various industrial processes. In the food industry, it is also used to protect wine from oxygen and bacteria [39].

Primary pollutants are injected into the atmosphere directly. Secondary pollutants are formed in the atmosphere through chemical and photochemical reactions from the primary pollutants [36].

All data from these six variables were normalized for the study. On the other hand, all of them are highly decorrelated. Table 1 shows the correlation matrix of the six pollutants of the case study.

It is worth mentioning that $O_3$ is the most independent pollutant, as its correlation coefficients with the rest of the variables are close to zero.

There are a total of 13,526 samples, as one sample per day (daily average) was collected for the twelve months of every year, between years 2000 and 2008, in the six stations analyzed in this study. Missing or corrupted data appear in all the variables in some rows, which are omitted for the study.

Table 2 shows the percentage of missing or corrupted data presented in each variable in the whole dataset.

All the samples with at least one missing or corrupted value were removed from the dataset.

## 4. Experiments, Results, and Discussion

The main target of this paper is to fill missing $O_3$ values in air pollution datasets. To do so, several imputation methods are comprehensively compared as described below.

*4.1. Experimental Settings.* The imputation methods described in Section 2 are applied to different datasets, all of them with the six variables described in Section 3:

(1) The Whole Dataset (WD), comprising the 13,526 samples: results for this datasets are shown in Section 4.2.

(2) The Season Dataset (SD): samples in WD are split in four subsets according to the four seasons of the year: spring (3,453 samples), summer (3,349 samples), autumn (3,295 samples), and winter (3,429 samples). Results for this dataset are shown in Section 4.3.

(3) The Type station Dataset (TD): samples in WD are split into two subsets according to the type of the station where the data come from; "urban" (6,763 samples) or "background" (6,763 samples). Results for this datasets are shown in Section 4.4.

For the three datasets, both statistical and neural imputation methods were applied and the performance is calculated through $n$-fold Cross-Validation (CV). The main idea behind CV is to split data, normally many times, for estimating the

Table 3: Linear regression and nonlinear regression results for the WD.

| Method | MSE | | Time (s) | |
|---|---|---|---|---|
| | Mean | STD | Mean | STD |
| MLR | $5.490E-06$ | $2.311E-08$ | **0.089** | 0.216 |
| MN-LR | **5.415E − 06** | $2.437E-08$ | 2.143 | 0.254 |

Table 4: Radial-basis function network results for the WD.

| # of neurons | MSE | | Time (s) | |
|---|---|---|---|---|
| | Mean | STD | Mean | STD |
| 10 | **5.104E − 06** | $2.723E-08$ | 0.050 | 1.091 |
| 30 | $5.108E-06$ | $1.273E-08$ | 0.050 | 1.091 |
| 50 | $5.105E-06$ | $2.513E-08$ | **0.047** | 0.098 |

risk, error, or performance of each algorithm. Part of data (the training samples) is used for training each algorithm, and the remaining part (the validation samples) is used for validating the algorithm(s). Then, CV selects the algorithm with the smallest estimated risk [41]. CV prevents from overfitting because the training sample is independent of the validation sample. The number of the $k$ parameters (data partitions) was 10 for all the experiments in the present study. It means that 90% of the data are used for training and 10% for validation. In the case of neural models, the training process is repeated ten times (one for each *fold*). In the case of MLP, training is also repeated for each training algorithm (see Section 2.2). For all the experiments the Mean and the Standard Deviation (STD) of the Mean Square Error (MSE) for the ten *folds* are presented in Tables 3–11. The Mean and the STD of the execution time (in seconds) are also presented in Tables 3–11 for the 10 *folds*.

For MLP and RBFN different network topologies have been applied: combinations of 10, 20, and 30 neurons in the hidden layer. Additionally, in the case of MLP, the model is trained 10 times with the same combination of parameters to reduce the effect of randomness and get more statistically significant results.

*4.2. Results from the Whole Dataset.* In this section, results in terms of MSE and execution time when applying MLR, MN-LR, RBFN, and MLP to the WD are presented.

In Tables 3 and 4, it can be observed that the MSE Mean values for the determination of the $O_3$ are very similar for the three applied methods (MLR, MN-LR, and RBFN). In the case of RBFN, slightly lower values of MSE are obtained, with the lowest one being obtained with 10 neurons in the hidden layer. Regarding execution times, the MN-LR method turns out to be the slowest and RBFN the quicker. The high values of STD for the runtime in the case of RBFN are due to the fact that it greatly varies from one fold to the others.

As it can be seen in Table 5, the LM, SCG, and BR training algorithms present the lowest values of MSE Mean in all cases (10, 30, and 50 neurons) and very close to those shown in Tables 3 and 4. The lowest value of MSE was obtained with the LM learning algorithm and 50 neurons. The learning algorithm that attained the worst results (in terms of MSE) is GDX. With respect to execution time, the SCG algorithm attained the best results, while LM and BR are the second best ones, while TB was the slowest of the five algorithms. Obviously, the training algorithms take more time when 50 neurons are defined in the hidden layer, the TB algorithm being the one with greatest effect.

*4.3. Results from the Season Dataset.* In Tables 6–8 results of applying MLR, MN-LR, RBFN, and MLP to subsets with data from the four seasons of the year (spring, summer, autumn, and winter) are presented.

In Tables 6 and 7 the 3 methods present similar values in MSE Mean, and the lowest MSE Mean is achieved by the RBFN with 50 neurons in the hidden layer for the summer season. The MSE Mean values are higher than that observed for the WD. The season of the year with the lowest values of MSE Mean is the summer. One reason may be that there are few variations in pollution conditions during summer time. This is due to the small variation in weather conditions during summer as well as low industrial activity and traffic in urban areas due to vacation time. Furthermore, correlation coefficients in more than 20 pollutants analyzed in [42] are higher for measurements in the summer compared with correlations for measurements over all days combined. The season of the year with the worst results in the calculation of the MSE has been the autumn in the case of the two regression techniques and RBFN, although the differences between the three seasons (spring, summer, and autumn) is not significant. In terms of execution time, it is probed once again that MN-LR is the slowest method, while RBFN is the quickest one, returning very similar results for the four seasons of the year.

In Table 8, similarly to Table 5, the training algorithms that achieve the best results in terms of MSE Mean are LM, SCG, and BR. LM achieved the best value of MSE Mean in 10 of the 12 cases shown in Table 8, being exceeded by BR by a minimum value for the winter and spring seasons with a configuration of 10 neurons. GDX records the worst MSE values in the 12 cases shown in Table 8. Again, the best MSE Mean is obtained for the summer season, reducing the MSE Mean in comparison with those registered by RBFN. The season of the year with the worst results in the calculation of

Table 5: Multilayer perceptron results for the WD.

| # of neurons | Training algorithm | MSE | | Time (s) | |
|---|---|---|---|---|---|
| | | Mean | STD | Mean | STD |
| 10 | LM | 4.731*E* − 06 | 5.143*E* − 08 | 0.070 | 0.287 |
| | GDX | 1.129*E* − 04 | 6.825*E* − 05 | 0.317 | 0.287 |
| | TB | 5.889*E* − 05 | 4.973*E* − 05 | 0.642 | 0.022 |
| | SCG | 5.216*E* − 06 | 1.514*E* − 07 | **0.060** | 0.001 |
| | BR | 4.775*E* − 06 | 1.092*E* − 07 | 0.074 | 0.003 |
| 30 | LM | 4.599*E* − 06 | 1.015*E* − 07 | 0.102 | 0.442 |
| | GDX | 4.045*E* − 04 | 3.523*E* − 04 | 0.481 | 0.442 |
| | TB | 4.223*E* − 05 | 2.087*E* − 05 | 1.420 | 0.025 |
| | SCG | 5.162*E* − 06 | 1.19*E* − 07 | 0.063 | 0.001 |
| | BR | 4.727*E* − 06 | 5.667*E* − 08 | 0.102 | 0.005 |
| 50 | LM | **4.512E − 06** | 8.952*E* − 08 | 0.160 | 1.080 |
| | GDX | 1.541*E* − 04 | 3.722*E* − 04 | 0.648 | 1.080 |
| | TB | 4.812*E* − 05 | 3.032*E* − 05 | 2.156 | 0.051 |
| | SCG | 5.014*E* − 06 | 8.322*E* − 08 | 0.068 | 0.001 |
| | BR | 4.731*E* − 06 | 1.099*E* − 07 | 0.161 | 0.010 |

Table 6: Linear regression and nonlinear regression results for the Season Dataset.

| Subset | Method | MSE | | Time (s) | |
|---|---|---|---|---|---|
| | | Mean | STD | Mean | STD |
| Spring | MLR | 1.895*E* − 05 | 1.406*E* − 07 | **0.085** | 0.208 |
| | MN-LR | 1.895*E* − 05 | 1.242*E* − 07 | 0.169 | 0.298 |
| Summer | MLR | 2.101*E* − 05 | 1.447*E* − 07 | **0.085** | 0.215 |
| | MN-LR | **1.343E − 05** | 1.365*E* − 07 | 0.665 | 0.321 |
| Autumn | MLR | 2.106*E* − 05 | 2.079*E* − 07 | **0.085** | 0.208 |
| | MN-LR | 2.101*E* − 05 | 1.447*E* − 07 | 0.677 | 0.259 |
| Winter | MLR | 1.895*E* − 05 | 1.406*E* − 07 | 0.088 | 0.214 |
| | MN-LR | 1.895*E* − 05 | 1.242*E* − 07 | 0.168 | 0.274 |

Table 7: Radial-basis function network results for the Season Dataset.

| Subset | # of neurons | MSE | | Time (s) | |
|---|---|---|---|---|---|
| | | Mean | STD | Mean | STD |
| Spring | 10 | 1.845*E* − 05 | 1.687*E* − 07 | 0.046 | 0.096 |
| | 30 | 1.847*E* − 05 | 1.06*E* − 07 | 0.050 | 0.098 |
| | 50 | 1.846*E* − 05 | 1.297*E* − 07 | 0.047 | 0.096 |
| Summer | 10 | 1.308*E* − 05 | 1.549*E* − 07 | **0.045** | 0.098 |
| | 30 | 1.310*E* − 05 | 1.389*E* − 07 | **0.045** | 0.096 |
| | 50 | **1.306E − 05** | 1.344*E* − 07 | 0.047 | 0.096 |
| Autumn | 10 | 1.986*E* − 05 | 1.176*E* − 07 | 0.046 | 0.096 |
| | 30 | 1.987*E* − 05 | 2.323*E* − 07 | 0.046 | 0.097 |
| | 50 | 1.987*E* − 05 | 2.199*E* − 07 | **0.045** | 0.095 |
| Winter | 10 | 1.845*E* − 05 | 1.687*E* − 07 | 0.046 | 0.100 |
| | 30 | 1.847*E* − 05 | 1.060*E* − 07 | **0.045** | 0.093 |
| | 50 | 1.846*E* − 05 | 1.297*E* − 07 | 0.046 | 0.096 |

Table 8: Multilayer perceptron results for the Season Dataset.

| Subset | # of neurons | Training algorithm | MSE | | Time (s) | |
|---|---|---|---|---|---|---|
| | | | Mean | STD | Mean | STD |
| Spring | 10 | LM | 1.696E − 05 | 2.550E − 07 | 0.063 | 0.086 |
| | | GDX | 3.298E − 04 | 8.39E − 04 | 0.183 | 0.086 |
| | | TB | 5.870E − 05 | 2.590E − 05 | 0.383 | 0.049 |
| | | SCG | 1.959E − 05 | 3.439E − 07 | 0.056 | 0.001 |
| | | BR | 1.695E − 05 | 1.988E − 07 | 0.076 | 0.015 |
| | 30 | LM | 1.531E − 05 | 4.107E − 07 | 0.068 | 0.217 |
| | | GDX | 6.652E − 04 | 1.051E − 03 | 0.210 | 0.217 |
| | | TB | 1.069E − 04 | 3.909E − 05 | 0.473 | 0.019 |
| | | SCG | 1.870E − 05 | 2.857E − 07 | 0.055 | 0.005 |
| | | BR | 1.562E − 05 | 3.915E − 07 | 0.071 | 0.225 |
| | 50 | LM | 1.473E − 05 | 4.138E − 07 | 0.080 | 0.181 |
| | | GDX | 6.286E − 04 | 1.200E − 03 | 0.253 | 0.181 |
| | | TB | 1.564E − 04 | 1.402E − 04 | 0.770 | 0.075 |
| | | SCG | 1.809E − 05 | 3.768E − 07 | 0.056 | 0.001 |
| | | BR | 1.580E − 05 | 3.697E − 07 | 0.089 | 0.490 |
| Summer | 10 | LM | 1.009E − 05 | 1.203E − 07 | 0.059 | 0.0663 |
| | | GDX | 4.718E − 04 | 4.791E − 04 | 0.166 | 0.0663 |
| | | TB | 7.65E − 05 | 6.231E − 05 | 0.355 | 0.0396 |
| | | SCG | 1.217E − 05 | 2.749E − 07 | **0.053** | 0.0007 |
| | | BR | 1.010E − 05 | 1.436E − 07 | 0.064 | 0.0059 |
| | 30 | LM | 9.171E − 06 | 2.713E − 07 | 0.065 | 0.440 |
| | | GDX | 8.070E − 04 | 1.446E − 03 | 0.202 | 0.440 |
| | | TB | 9.117E − 05 | 5.051E − 05 | 0.500 | 0.047 |
| | | SCG | 1.118E − 05 | 4.118E − 07 | 0.056 | 0.001 |
| | | BR | 9.822E − 06 | 3.107E − 07 | 0.073 | 0.004 |
| | 50 | LM | **8.673E − 06** | 3.284E − 07 | 0.083 | 0.225 |
| | | GDX | 4.572E − 04 | 9.864E − 04 | 0.253 | 0.225 |
| | | TB | 1.269E − 04 | 6.38E − 05 | 0.743 | 0.019 |
| | | SCG | 1.089E − 05 | 1.561E − 07 | 0.057 | 0.001 |
| | | BR | 9.851E − 06 | 2.165E − 07 | 0.085 | 0.003 |
| Autumn | 10 | LM | 1.622E − 05 | 2.146E − 07 | 0.061 | 0.101 |
| | | GDX | 1.598E − 04 | 3.096E − 04 | 0.168 | 0.101 |
| | | TB | 7.248E − 05 | 3.589E − 05 | 0.351 | 0.058 |
| | | SCG | 1.904E − 05 | 2.617E − 07 | 0.055 | 0.001 |
| | | BR | 1.628E − 05 | 7.680E − 07 | 0.071 | 0.009 |
| | 30 | LM | 1.495E − 05 | 3.564E − 07 | 0.067 | 0.196 |
| | | GDX | 1.045E − 03 | 1.520E − 03 | 0.204 | 0.196 |
| | | TB | 1.09E − 04 | 6.737E − 05 | 0.506 | 0.048 |
| | | SCG | 1.808E − 05 | 2.756E − 07 | 0.054 | 0.001 |
| | | BR | 1.522E − 05 | 3.456E − 07 | 0.069 | 0.001 |
| | 50 | LM | 1.401E − 05 | 1.926E − 06 | 0.079 | 0.307 |
| | | GDX | 5.676E − 04 | 1.700E − 03 | 0.240 | 0.307 |
| | | TB | 1.005E − 04 | 6.447E − 05 | 0.734 | 0.029 |
| | | SCG | 1.758E − 05 | 5.509E − 07 | 0.054 | 0.001 |
| | | BR | 1.559E − 05 | 6.591E − 07 | 0.083 | 0.103 |

Table 8: Continued.

| Subset | # of neurons | Training algorithm | MSE | | Time (s) | |
|---|---|---|---|---|---|---|
| | | | Mean | STD | Mean | STD |
| Winter | 10 | LM | 1.573$E$ − 05 | 2.204$E$ − 07 | 0.060 | 0.105 |
| | | GDX | 3.380$E$ − 04 | 8.630$E$ − 04 | 0.170 | 0.105 |
| | | TB | 9.792$E$ − 05 | 1.398$E$ − 04 | 0.356 | 0.008 |
| | | SCG | 1.811$E$ − 05 | 3.452$E$ − 07 | 0.055 | 0.001 |
| | | BR | 1.558$E$ − 05 | 1.229$E$ − 06 | 0.065 | 0.016 |
| | 30 | LM | 1.423$E$ − 05 | 2.490$E$ − 07 | 0.070 | 0.122 |
| | | GDX | 5.721$E$ − 04 | 7.005$E$ − 04 | 0.215 | 0.122 |
| | | TB | 1.253$E$ − 04 | 4.611$E$ − 05 | 0.490 | 0.035 |
| | | SCG | 1.710$E$ − 05 | 3.509$E$ − 07 | 0.056 | 0.001 |
| | | BR | 1.475$E$ − 05 | 2.85$E$ − 07 | 0.103 | 0.096 |
| | 50 | LM | 1.355$E$ − 05 | 6.746$E$ − 07 | 0.080 | 0.155 |
| | | GDX | 6.788$E$ − 04 | 5.702$E$ − 04 | 0.247 | 0.155 |
| | | TB | 1.324$E$ − 04 | 1.253$E$ − 04 | 0.757 | 0.011 |
| | | SCG | 1.670$E$ − 05 | 4.094$E$ − 07 | 0.055 | 0.001 |
| | | BR | 1.478$E$ − 05 | 2.93$E$ − 07 | 0.080 | 0.115 |

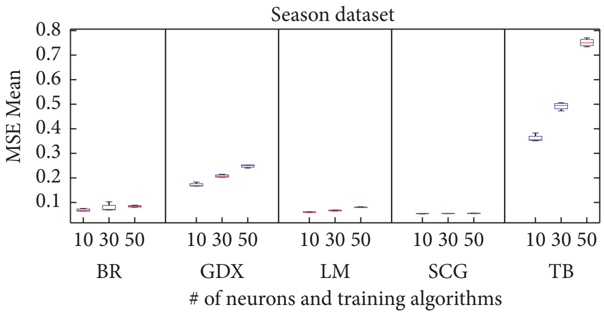

Figure 2: Boxplot for the MLP applied to the Season Dataset (SD).

the MSE has been the spring, although the difference in this term between spring, autumn, and winter is minimal.

In terms of execution time, it can be said that the SCG algorithm is the fastest one (in terms of mean execution time), with slight variations (low STD). LM and BR perform very well according to runtime with very similar result. Finally, the TB algorithm is the slowest one in the 12 cases shown in Table 8. This algorithm is very sensitive in its execution time to the increase in the number of neurons in the hidden layer. It is worth mentioning that the best value in terms of execution time has been obtained by SCG and for the summer season, the same for which the best value of MSE is achieved.

Figure 2 shows the boxplot for the results shown in Table 8. Each box represents the MSE Mean values for the whole dataset (four seasons), for a certain number of neurons and a training algorithm.

In Figure 2 it can be observed that the LM and SCG training algorithms outperform the other algorithms and that the TB algorithm achieved the worst results. It is also worth mentioning that, in general terms, increasing the number of neurons in the hidden layer causes an increase in the MSE due to the loss of generalization capability of the models (especially in the TB and GDX algorithms). The difference between the 25th and 75th percentiles is also higher in the case of algorithms achieving poor results in the Season Dataset, especially for the TB training algorithm.

### *4.4. Results from the Station Type Dataset.*

Tables 9–11 show the results of applying the four techniques to two different subsets, according to the station type: urban or background (see Section 4.1 for further details).

MLR, according to Sections 4.2 and 4.3, again MN-LR, and RBF achieve similar results in estimating the MSE, but in this occasion it is higher than in Sections 4.2 and 4.3. In turn, the urban stations get a better MSE than the background stations; this indicates that the pollution levels are more constant in the urban stations than in the ones furthest from the center. In terms of the execution time, the MN-LR is again the slowest method. The RBFN shows a more efficient response than the regression methods.

For MLP, and in similar way compared to the other datasets (Sections 4.2 and 4.3), the training algorithms which achieved the lowest MSE Mean values are LM, SCG, and BR. These values are similar to those obtained by RBFN (Table 10) and lower than the values associated with the regression techniques in Table 9. According to the station type, generally speaking, lower MSE values were obtained for the “urban” stations, in comparison with “background” stations. The lowest MSE value was obtained for “urban” stations with 50 neurons and LM algorithm. The only training algorithm which returns higher values of MSE Mean for the “urban” stations than for the “background” stations is GDX. This happened for the three different numbers of neurons, while the other four algorithms get lower values of MSE for the “urban” stations for the different numbers of neurons in the hidden layer. The lower value in the MSE makes the “urban” stations easier to estimate the missing $O_3$ values; this is due to fewer variations in the pollution values in

Table 9: Linear regression and nonlinear regression results for the Station Type Dataset.

| Subset | Method | MSE |  | Time (s) |  |
|---|---|---|---|---|---|
|  |  | Mean | STD | Mean | STD |
| Urban | MLR | $9.922E-06$ | $3.297E-08$ | **0.053** | 0.190 |
|  | MN-LR | $\mathbf{9.352E-06}$ | $3.840E-08$ | 1.172 | 0.249 |
| Background | MLR | $1.148E-05$ | $2.495E-08$ | 0.092 | 0.214 |
|  | MN-LR | $1.141E-05$ | $6.084E-08$ | 0.204 | 0.297 |

Table 10: Radial-basis function network results for the Station Type Dataset.

| Subset | # Neurons | MSE |  | Time (s) |  |
|---|---|---|---|---|---|
|  |  | Mean | STD | Mean | STD |
| Urban | 10 | $8.480E-06$ | $4.001E-08$ | **0.047** | 0.086 |
|  | 30 | $8.486E-06$ | $3.531E-08$ | **0.047** | 0.094 |
|  | 50 | $\mathbf{8.477E-06}$ | $5.212E-08$ | **0.047** | 0.083 |
| Background | 10 | $8.559E-06$ | $4.078E-08$ | 0.047 | 0.095 |
|  | 30 | $9.654E-06$ | $3.401E-08$ | 0.047 | 0.096 |
|  | 50 | $1.141E-05$ | $6.084E-08$ | 0.045 | 0.093 |

the predictive variables throughout the year in this type of stations.

In terms of execution time, the SCG algorithm is the quickest one in the six cases shown in Table 11 followed by LM, with no big difference depending on the number of neurons in the hidden layer, only a little faster with 10 neurons. The slowest train algorithm is again TB in the six cases, as it was identified from these results with the exposed results in Tables 5 and 8.

*4.5. Discussion.* The two applied regression techniques (MLR and MN-LR) obtained similar values of MSE in most cases, in terms of both Mean and STD. However, MN-LR obtained poor results according to execution time (Tables 3, 6, and 9), even worse than the slowest training algorithm for MLP (GDX and TB in Tables 5, 8, and 11).

For the ANN models (RBFN and MLP), different combinations of neurons in the hidden layer were compared. For the sake of brevity, only the results for 10, 30, and 50 neurons (Tables 4, 5, 7, 8, 10, and 11) have been included in the present paper. In the case of RBFN, the best execution times are achieved, outperforming the fastest algorithm for MLP (SCG in Tables 5, 8, and 11). In the case of MLP, varying results have been obtained, depending on the training algorithm applied, obtaining the best results (in terms of MSE) when learning through the LM and SCG algorithms. SCG algorithm additionally is the fastest one. GDX has been identified as the algorithm with worst error, as can be seen in Tables 5, 8, and 11. No significant improvement is observed in the estimation of missing values according to MSE when increasing the number of neurons of the hidden layer. On the contrary, the selection of the training algorithm has been identified as a key factor when applying MLP. An increase in the number of neurons in RBFN does not affect considerably the accuracy of the results in terms of MSE and execution time (see Tables 4, 7, and 10). MLP achieved a better value of MSE if the training algorithm is properly selected.

Taking into account the different datasets, the lowest MSE for the Season Dataset is obtained for the summer season when applying the LM training algorithm with 50 neurons, without big differences between the other three seasons of the year. For the spring, autumn, and winter seasons the best MSE corresponds to the LM algorithm combined with 50 neurons. In terms of execution time, the fastest experiment was that applying RBF with 10 and 30 neurons for the summer season, SCG for 50 neurons in the case of spring season, RBF with 50 neurons for autumn season, and RBF with 30 neurons for the winter season. In the case of the Station Type Dataset, "urban" stations, the best results in terms of MSE for the "urban" stations and for the summer season are accompanied by the lowest execution times. It must be mentioned that good results have been obtained, in terms of MSE, when applying the four imputation methods to the WD. This fact indicates no great variations neither between the weather seasons of the year nor between the analyzed types of station ("urban" and "background").

## 5. Conclusions

In the present work, several different imputation methods are proposed for dealing with missing $O_3$ values in multidimensional real-life datasets with air-quality information. To do this, two multiple-regression techniques (linear and nonlinear) and two ANN models (RBFN and MLP) with different training algorithms and different number of neurons in the hidden layer have been compared. As a validation scheme, 10-fold cross-validation has been applied to the different datasets. The imputation task has been carried out firstly on the complete dataset, and on different datasets, where the original data are split according to two criteria: according to the season and according to the station type.

Table 11: Multilayer perceptron results for the Station Type Dataset.

| Subset | # of neurons | Training algorithm | MSE | | Time (s) | |
|---|---|---|---|---|---|---|
| | | | Mean | STD | Mean | STD |
| Urban | 10 | LM | $6.785E-06$ | $2.563E-07$ | 0.062 | 0.092 |
| | | GDX | $8.761E-05$ | $1.457E-04$ | 0.223 | 0.092 |
| | | TB | $5.551E-05$ | $2.967E-05$ | 0.411 | 0.029 |
| | | SCG | $8.305E-06$ | $3.095E-07$ | 0.058 | 0.001 |
| | | BR | $6.572E-06$ | $2.574E-07$ | 0.069 | 0.002 |
| | 30 | LM | $6.237E-06$ | $1.037E-07$ | 0.081 | 0.176 |
| | | GDX | $2.256E-04$ | $4.360E-04$ | 0.292 | 0.176 |
| | | TB | $6.583E-05$ | $2.762E-05$ | 0.834 | 0.028 |
| | | SCG | $7.752E-06$ | $1.808E-07$ | 0.058 | 0.001 |
| | | BR | $6.527E-06$ | $4.274E-07$ | 0.086 | 0.004 |
| | 50 | LM | $\mathbf{6.059E-06}$ | $\mathbf{7.950}E-08$ | 0.110 | 0.230 |
| | | GDX | $6.550E-04$ | $1.165E-04$ | 0.379 | 0.230 |
| | | TB | $6.975E-05$ | $2.578E-05$ | 1.168 | 0.026 |
| | | SCG | $7.527E-06$ | $1.559E-07$ | 0.062 | 0.002 |
| | | BR | $6.437E-06$ | $1.117E-07$ | 0.108 | 0.014 |
| Background | 10 | LM | $6.964E-06$ | $7.935E-08$ | 0.064 | 0.125 |
| | | GDX | $1.125E-04$ | $2.734E-04$ | 0.226 | 0.125 |
| | | TB | $1.109E-04$ | $7.399E-05$ | 0.444 | 0.069 |
| | | SCG | $8.106E-06$ | $2.258E-07$ | **0.056** | 0.002 |
| | | BR | $7.036E-06$ | $1.47E-07$ | 0.069 | 0.008 |
| | 30 | LM | $6.469E-06$ | $6.612E-08$ | 0.077 | 0.204 |
| | | GDX | $1.9E-04$ | $2.6E-04$ | 0.294 | 0.204 |
| | | TB | $8.202E-05$ | $2.319E-05$ | 0.848 | 0.012 |
| | | SCG | $7.707E-06$ | $1.100E-07$ | 0.059 | 0.001 |
| | | BR | $6.813E-06$ | $3.061E-07$ | 0.079 | 0.010 |
| | 50 | LM | $6.184E-06$ | $6.764E-08$ | 0.107 | 0.317 |
| | | GDX | $3.1E-04$ | $8.9E-04$ | 0.383 | 0.317 |
| | | TB | $7.926E-05$ | $5.183E-05$ | 1.205 | 0.070 |
| | | SCG | $7.625E-06$ | $9.986E-08$ | 0.062 | 0.001 |
| | | BR | $6.743E-06$ | $1.253E-07$ | 0.111 | 0.006 |

The following conclusions are worth mentioning:

(1) MLR and MN-LR attained very similar results in terms of MSE and execution time. These results are slightly worse than those obtained by the two ANN models (RBFN and MLP). The lowest value of MSE has been obtained for the WD (applying MN-LR technique) and the highest one for the SD (also applying MN-LR technique).

(2) In the case of RBFN, slight differences have been obtained when varying the number of neurons in the hidden layer, in terms of both the MSE and the execution time. The best results have been obtained for the WD (with 10 neurons in the hidden layer) and the worst for the SD (with 50 neurons in the hidden layer), as it happened for MLR and MN-LR.

(3) In the case of MLP, the best results are achieved when using the LM training algorithm and a number of 50 neurons in the hidden layer. As in the previous case (RBFN), the best results are obtained for the WD and the worst results for the SD, with small differences between the results in the three datasets. These are the best result from the whole experimentation in the present paper. The results obtained by MLP improve those obtained by RBFN, only when applying the LM training algorithm.

(4) The CV technique guarantees reliability of the results when dealing with large datasets.

As future work, the application of additional artificial-intelligence models for the imputation of $O_3$ and other pollutants is proposed, comparing the results with those obtained in the present study.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## References

[1] The Pubchem Project, "Ozone," https://www.ncbi.nlm.nih.gov/pubmed/.

[2] W. L. Chameides, F. Fehsenfeld, and M. O. Rodgers, "Ozone precursor relationships in the ambient atmosphere," *Journal of Geophysical Research: Atmospheres*, vol. 97, no. 5, pp. 6037–6055, 1992.

[3] United States Environmental Protection Agency, "What is Ozone?" https://www.epa.gov/ozone-pollution-and-your-patients-health/what-ozone.

[4] United States Environmental Protection Agenc, "Health and Environmental Effects of Ozone Layer Depletion," https://www.epa.gov/ozone-layer-protection/health-and-environmental-effects-ozone-layer-depletion.

[5] European Union Law, "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," 2008, http://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:32008L0050.

[6] C. C. Turrado, M. D. C. M. López, F. S. Lasheras, B. A. R. Gómez, J. L. C. Rollé, and F. J. D. C. Juez, "Missing data imputation of solar radiation data under different atmospheric conditions," *Sensors*, vol. 14, no. 11, pp. 20382–20399, 2014.

[7] P. Schmitt, J. S. Mandel, and M. Guedj, "A Comparison of Six Methods for Missing Data Imputation," *Journal of Biometrics Biostatistics*, p. 6, 2015.

[8] T. D. Pigott, "A Review of Methods for Missing Data," *Educational Research and Evaluation*, vol. 7, no. 4, pp. 353–383, 2001.

[9] Á. Arroyo et al., "A hybrid intelligent system for the analysis of atmospheric pollution: a case study in two European regions," *Logic Journal of the IGPL*, vol. 25, no. 6, pp. 915–937, 2017.

[10] J. L. Schafer, "Multiple imputation: a primer," *Statistical Methods in Medical Research*, vol. 8, no. 1, pp. 3–15, 1999.

[11] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen, "Methods for imputation of missing values in air quality data sets," *Atmospheric Environment*, vol. 38, no. 18, pp. 2895–2907, 2004.

[12] A. Plaia and A. L. Bondì, "Single imputation method of missing values in environmental pollution data sets," *Atmospheric Environment*, vol. 40, no. 38, pp. 7316–7330, 2006.

[13] M. P. Gómez-Carracedo, J. M. Andrade, P. López-Mahía, S. Muniategui, and D. Prada, "A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets," *Chemometrics and Intelligent Laboratory Systems*, vol. 134, pp. 23–33, 2014.

[14] J. L. Schafer, *Analysis of Incomplete Multivariate Data*, Hall/CRC Monographs on Statistics & Applied Probability, Chapman & Hall, New York, NY, USA, 1997.

[15] S. Ghorbani and M. C. Desmarais, "Performance Comparison of Recent Imputation Methods for Classification Tasks over Binary Data," *Applied Artificial Intelligence*, vol. 31, no. 1, pp. 1–22, 2017.

[16] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, 2010.

[17] M. D. Julie and B. Kannan, "Attribute reduction and missing value imputing with ANN: Prediction of learning disabilities," *Neural Computing and Applications*, vol. 21, no. 7, pp. 1757–1763, 2012.

[18] M. S. Gashler, M. R. Smith, R. Morris, and T. Martinez, "Missing value imputation with unsupervised backpropagation," *Computational Intelligence*, vol. 32, no. 2, pp. 196–215, 2016.

[19] P. Coulibaly and N. D. Evora, "Comparison of neural network methods for infilling missing daily weather records," *Journal of Hydrology*, vol. 341, no. 1-2, pp. 27–41, 2007.

[20] I. A. Gheyas and L. S. Smith, "A neural network-based framework for the reconstruction of incomplete data sets," *Neurocomputing*, vol. 73, no. 16-18, pp. 3039–3065, 2010.

[21] E.-L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, and M.-D. Cubiles-de-la-Vega, "Missing value imputation on missing completely at random data using multilayer perceptrons," *Neural Networks*, vol. 24, no. 1, pp. 121–129, 2011.

[22] R. R. Andridge and R. J. A. Little, "A review of hot deck imputation for survey non-response," *International Statistical Review*, vol. 78, no. 1, pp. 40–64, 2010.

[23] D. Sovilj, E. Eirola, Y. Miche et al., "Extreme learning machine for missing data using multiple imputations," *Neurocomputing*, vol. 174, pp. 220–231, 2016.

[24] E.-L. Silva-Ramírez, R. Pino-Mejías, and M. López-Coello, "Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns," *Applied Soft Computing*, vol. 29, pp. 65–74, 2015.

[25] S. Azim and S. Aggarwal, "Hybrid model for data imputation: using fuzzy c means and multi layer perceptron," in *Proceedings of the 4th IEEE International Advance Computing Conference (IACC '14)*, pp. 1281–1285, Gurgaon, India, February 2014.

[26] University of Yale, "Linear Regression," http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm.

[27] K. Pearson and A. Lee, "On the generalised probable error in multiple normal correlation," *Biometrika*, vol. 6, no. 1, pp. 59–68, 1908.

[28] University of Yale, "Multiple Linear Regression," 2017, http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm.

[29] M. H. Kutner, C. J. Nachtsheim, and J. Neter, *Applied Linear Regression Models*, McGraw Hill.

[30] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '89)*, vol. 1, pp. 593–605, Washington, DC, USA, June 1989.

[31] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric Environment*, vol. 32, no. 14-15, pp. 2627–2636, 1998.

[32] C. C. Yu and B. D. Liu, "A backpropagation algorithm with adaptive learning rate and momentum coefficient," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2002.

[33] R. P. Lippmann, "Pattern Classification Using Neural Networks," *IEEE Communications Magazine*, vol. 27, no. 11, pp. 47–50, 1989.

[34] J. Park and I. W. Sandberg, "Universal approximation using radial basis function networks," *Neural Computation*, vol. 3, no. 2, pp. 246–257, 1991.

[35] European Commission, "Castilla y León," 2016, https://ec.europa.eu/growth/tools-databases/regional-innovation-monitor/base-profile/castilla-y-le%C3%B3n-0.

[36] Castilla y León Government, “Castilla y León Air Quality network,” https://medioambiente.jcyl.es/web/jcyl/MedioAmbiente/es/Plantilla66y33//_/_/_.

[37] Government of spain, “Open Data,” http://datos.gob.es/.

[38] Castilla y León Government, “Castilla y León Open Data,” https://datosabiertos.jcyl.es/web/jcyl/set/es/medio-ambiente/calidad_aire_estaciones/1284212701893.

[39] National Center for Biotechnology Information, “PubChem Compound,” 2017, https://www.ncbi.nlm.nih.gov/pubmed/.

[40] Dekati, *ISO 23210 Measurements of PM10 and PM2.5*, 2017, https://www.dekati.com/applications/stationary-source-emissions/emission-monitoring/pm10-and-pm25-according-to-iso23210.

[41] S. Arlot and A. Celisse, “A survey of cross-validation procedures for model selection,” *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[42] I. Levy, C. Mihele, G. Lu, J. Narayan, and J. R. Brook, “Evaluating multipollutant exposure and urban air quality: Pollutant interrelationships, neighborhood variability, and nitrogen dioxide as a proxy pollutant,” *Environmental Health Perspectives*, vol. 122, no. 1, pp. 65–72, 2014.