© by PSP Volume 28 – No. 2/2019 pages 681-686 Fresenius Environmental Bulletin

# ESTIMATION OF $NO_2$ CONCENTRATION VALUES IN A MONITORING SENSOR NETWORK USING A FUSION APPROACH

**Javier Gonzalez-Enrique[1,*], Ignacio J Turias[1], Juan Jesus Ruiz-Aguilar[2], Jose Antonio Moscoso-Lopez[2], Jose Jerez-Aragones[3], Leonardo Franco[3]**

[1]Department of Computer Science Engineering, Polytechnic School of Engineering, University of Cádiz, Algeciras, Spain.
[2]Department of Industrial and Civil Engineering, Polytechnic School of Engineering, University of Cádiz, Algeciras, Spain.
[3]Department of Computer Science, ETS Computer Science, University of Málaga, Spain.

## ABSTRACT

This study is focused on calculation of a reliable estimation of the hourly concentration value of $NO_2$ at a monitoring station based on a data fusion approach. Different feature selection procedures have been tested and their results were used as inputs to an artificial neural network (ANN) two-stage approach. The final aim is to develop a data fusion estimation tool in order to aid the decision-making in the monitoring process. ANN models were trained using backpropagation and early stopping in order to avoid overfitting. Furthermore, the study compares the different combinations of methods to estimate hourly $NO_2$ concentration values. The comparison was made using different performance indexes (R, d, MAE, MSE). The data from the Algeciras Bay was used as a case study. This approach may become a supporting tool for different purposes such as missing data imputation, automatic detection of decalibration or imprecise data in monitoring networks.

**KEYWORDS:**
Artificial neural networks, data fusion, atmospheric pollution.

## INTRODUCTION

Air pollution has become an important environmental problem in metropolitan areas [1]. EU and many national environmental agencies define quality objectives for allowable levels of atmospheric pollutants [2]. An accurate air pollutants monitoring is required for air quality management to provide proper actions and control strategies [3]. The main urban-related pollutants are CO, $NO_x$, hydrocarbons, and particles. Nitrogen dioxide ($NO_2$) it is a very reactive toxic gas, which produces a strong odour, is highly corrosive and have great irritating power. $NO_2$ is the main responsible in the formation of the smog and acid rain in urban areas. In turn, it produces acute and chronic effects, particularly in sensitive people. The control and reduction of the high levels of $NO_2$ are one of the main targets of governments.

Many atmospheric air quality reports have been regularly published in recent years. The meteorological conditions have a marked effect over the pollutants concentration as they diminish the ability of the atmosphere to disperse pollutants. The relationships between meteorological conditions and air pollutions may suggest estimating concentration values using a multivariate regression model. These relationships (pollution-weather) are very complex and may have nonlinear properties. Thus, better performance can be obtained using Artificial Neural Networks (ANNs) [4]. ANNs have become an alternative to traditional methods and they are an important tool to model air pollution with a wide range of pollutants at several time scales [5]. ANN approaches have been frequently used in atmospheric and air quality modelling studies [6]. Some studies have used ANNs to estimate air pollution peaks using meteorological variables [7].

Environmental monitoring is a very important task. An environmental monitoring network consists of a number of sensor nodes (few tens to thousands) working together to monitor a region to obtain data about the atmospheric pollution. Typically, network maintenance such as detecting failures is difficult and expensive since there are so many nodes. It is necessary to make these applications more reliable and robust in the real world. One interesting approach deals with the data fusion of the multiple sensors at the monitoring stations. Thus, sensor nodes can be able to self-organize and identify problems themselves. Data fusion has several advantages [8], mainly involving enhancements in data authenticity or availability. From network management and reliability perspectives, it is important that sensor nodes are capable of self-organizing themselves. Soft computing techniques have received an increasing interest in research on multi-sensor data fusion technology [9].

The main objective of the present work is to determine the concentration values of Nitrogen Dioxide ($NO_2$) at a certain monitoring station using a data

fusion approach of the meteorological variables, measured in other locations, and the concentration values measured in other monitoring stations located in the area of study. In this paper, authors have used different feature selection approaches in order to provide the best features as inputs to the following estimation stage. Different topologies of ANNs have been used in order to estimate the concentration values at a monitoring station as a function of the selected features. The different approaches have been compared using an experimental framework. Authors have also used a previously designed resampling procedure [10, 11] in order to statistically analyze and compare the results.

The rest of this paper is organized as follows: Section 2 describes the data and the region. Section 3 gives an overview of the feature selection methods. Section 4 gives a quick description of the artificial neural networks design. Section 5 presents the experimental framework. Section 6 discusses the results. Finally, Section 7 concludes the paper.

## DATA AND AREA DESCRIPTION

This work is located in the Bay of Algeciras region, which is one of the most industrialized areas in Andalusia (South of Spain). Besides, Algeciras has the most important port of the Mediterranean Sea and a freeway (A-7), which links the entire region with almost 70,000 passenger cars and 4,000 heavy vehicles per day. It is a very populated region, with almost 300,000 people living in different towns, and many sources of particulate and gaseous air pollution are present. Despite that, there are only a few studies devoted to the study of air pollution in the region.

A monitoring network located in the area of study has supplied the database used in this work. This network is composed of fourteen monitoring stations. Additionally, five weather stations have supplied meteorological data. On the one hand, the monitoring stations have recorded an hourly database of $NO_2$ concentrations during a period of six years (2010-2015). These measures are controlled by the Environmental Agency of the Regional Andalusian Government (Spain). On the other hand, the meteorological information has been extracted from the weather stations located at three different sites, including a meteorological tower where the variables (wind speed and wind direction) have been measured at 60-meter height. No imputation methods have been used. Table 1 shows the stations that make up the monitoring network. Codes 1 to 14 indicate $NO_2$ monitoring stations, while codes W1 to W5 indicate weather stations. Figure 1 shows each station represented by its code. A description of each recorded variable can be found in Table 2.

**TABLE 1**
**Monitoring and meteorological stations.**

| Code | Station name |
|---|---|
| 1 | EPSA Algeciras |
| 2 | Campamento |
| 3 | Los Cortillijos |
| 4 | Esc. Hostelería |
| 5 | Col. Los Barrios |
| 6 | Col. Carteya |
| 7 | El Rinconcillo |
| 8 | Palmones |
| 9 | San Roque |
| 10 | El Zabal |
| 11 | Economato |
| 12 | Guadarranque |
| 13 | La Línea |
| 14 | Madrevieja |
| W1 | La Línea weather station |
| W2 | Los Barrios weather station |
| W3 | CEPSA weather station (10 meters) |
| W4 | CEPSA weather station (60 meters) |
| W5 | CEPSA weather station (15 meters) |


**FIGURE 1**
**Location of the monitoring and weather stations. Map data: Google, Data SIO, NOAA, U.S. Navy, NGA, GEBCO TerraMetrics.**

## FEATURE SELECTION METHODS

Two main approaches can be considered when the objective is to reduce data dimensionality: feature selection and feature transformation.

On the one hand, feature selection has been widely studied in the last years in the literature [12]. Most researchers agree that there is not a best general method. Thus, a good method has to be found out in each particular problem. The performance of the feature selection methods relies on the performance of the learning method used afterwards and it can vary notably [13]. Therefore, the majority of the interesting comparative studies are focused on the problem to be solved [14]. Feature selection methods can be divided into three categories: filter, wrapper and embedded methods [15]. Among them, filter methods are based on the suppression of the least significant features.

**TABLE 2**
**Variable description.**

| Variable number | Meaning | Variable number | Meaning |
|---|---|---|---|
| 1 | EPSA Algeciras - $NO_2$ ($mg/m^3$) | 20 | W2 - Wind direction (degrees) |
| 2 | Campamento - $NO_2$ ($mg/m^3$) | 21 | W2 - Relative humidity (%) |
| 3 | Los Cortillijos - $NO_2$ ($mg/m^3$) | 22 | W2 - Rainfall ($l/m^2$) |
| 4 | Esc. Hostelería - $NO_2$ ($mg/m^3$) | 23 | W2 - Atmospheric pressure (hPa) |
| 5 | Col. Los Barrios - $NO_2$ ($mg/m^3$) | 24 | W2 - Solar radiation ($w/m^2$) |
| 6 | Col. Carteya - $NO_2$ ($mg/m^3$) | 25 | W3 - Relative humidity (%) |
| 7 | El Rinconcillo - $NO_2$ ($mg/m^3$) | 26 | W3 - Rainfall ($l/m^2$) |
| 8 | Palmones - $NO_2$ ($mg/m^3$) | 27 | W3 - Atmospheric pressure (hPa) |
| 9 | San Roque - $NO_2$ ($mg/m^3$) | 28 | W3 - Solar radiation ($w/m^2$) |
| 10 | El Zabal - $NO_2$ ($mg/m^3$) | 29 | W4 - Wind direction (degrees) |
| 11 | Economato - $NO_2$ ($mg/m^3$) | 30 | W4 - Temperature (C) |
| 12 | Guadarranque - $NO_2$ ($mg/m^3$) | 31 | W4 - Wind speed (km/h) |
| 13 | La Línea - $NO_2$ ($mg/m^3$) | 32 | W5 - Wind direction (degrees) |
| 14 | Madrevieja - $NO_2$ ($mg/m^3$) | 33 | W5 - Relative humidity (%) |
| 15 | W1 - Wind direction (degrees) | 34 | W5 - Rainfall ($l/m^2$) |
| 16 | W1 - Relative humidity (%) | 35 | W5 - Atmospheric pressure (hPa) |
| 17 | W1 - Rainfall ($l/m^2$) | 36 | W5 - Solar radiation ($w/m^2$) |
| 18 | W1 - Temperature (C) | 37 | W5 - Temperature (C) |
| 19 | W1 - Wind speed (km/h) | 38 | W5 - Wind speed (km/h) |

On the other hand, feature transformation methods are based on the transformation of the existing features into a lower dimensional space.

In this work, two different approaches have been applied: feature selection using filter methods and feature transformation. Those approaches rely on the general characteristics of training data and carry out the dimensionality reduction process as a pre-processing step with independence of the learning algorithm. Specifically, the following methods have been used in this work:

1. **Feature selection using filter methods:**
- Regression p-values: The p-value for each term in a regression analysis tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.
- ReliefF [16]: It works by randomly sampling an instance from the data and then locating its nearest neighbour from the same and opposite class. The values of the attributes of the nearest neighbours are compared to the sampled instance and used to update relevance scores for each attribute.

2. **Feature transformation methods:**
- Principal Component Analysis (PCA): A well-known orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [17]. The first principal component has the largest variance. The new vectors are an uncorrelated orthogonal basis set.
- Non-negative matrix factorization (NMF): It is a dimension-reduction technique [18] based on a low-rank approximation of the feature space. Besides providing a reduction in the number of features. NMF finds nonnegative matrices W and H, respectively, that minimize the norm of the difference X – WH. W and H are thus approximate nonnegative factors of X, and the k columns of W represent transformations of the variables in X.

## ARTIFICIAL NEURAL NETWORKS

Multi-Layer Perceptron (MLPs) are considered here since they have proven their efficiency to describe the nonlinear relationships between a set of input and output values. For these reasons, MLP is the most widely used neural network architecture for regression problems where all the neurons are organized in a layered feedforward topology. Feedforward ANNs using backpropagation [19] are proved to be universal approximators [20]. However, there is no a general way to determine the best ANN to solve a problem. The aim is to achieve the best generalization capability (for a test set of samples) avoiding overfitting. In order to do that, an early stopping has been used together with a resampling procedure using a cross-validation scheme [21].

## EXPERIMENTAL DESIGN

The objective of this study is to develop a model to estimate the hourly concentration value of $NO_2$ at the EPSA Algeciras monitoring station (see Table 1), according to the values of the rest of stations and the meteorological variables measured in the area of study. Thus, a database of 37 variables was used to obtain this estimation (see Table 2).

The use of soft computing techniques in engineering applications has increased over the last years [22]. In this context, the choice of the evaluation

technique is critical. In order to evaluate the performance of the model, cross-validation was used to determine when overfitting starts during supervised training of a neural network. This technique improves the performance and robustness for prediction applications as it has been proven empirically in a large number of works [23].

The values of the weights and bias were adjusted according to the Levenberg-Marquardt optimization algorithm. The original data set was preprocessed (only complete records have been used) and all the models were defined using a different number of hidden neurons (nhiddens) and trained using the 2-fold cross-validation (2-CV) as validation technique. This validation procedure was repeated 20 times, taking the average value of all of them, so that the randomness of the process was guaranteed. In order to select the optimal structure of the model, four statistical performance indexes have been measured: the standard correlation coefficient (R), the index of agreement (d) [24], the mean square error (MSE) and the mean absolute error (MAE). This resampling procedure has been applied in the proposed two-stage approach.

In the first stage, the different feature selection methods were tested using ANNs as an induction method, and the best models were selected. The four feature selection methods were applied to the original data set. In the case of regression p-values method, variables (see Table 2) were sorted according to their ascending p-values. For ReliefF Algorithm, variables were sorted according to their descending weights. In the case of PCA, a new space was obtained and the components were ordered by the proportion of variance explained. Finally, different approximations were obtained for NMF, which were ordered by their ranks. After that, the best possible model was obtained for every of the aforementioned selection methods using ANNs with a stepwise [25] inspired procedure. The approach was the same in the case of regression p-values, ReliefF and PCA, but had slight differences in the case of NMF:

- **Regression p-values, ReliefF and PCA:** Starting with a dataset composed only of the very best variable (or first component in the case of PCA), ANNs models were developed and their performance indicators were obtained. In a second step, the next variable according to the sort order (or component for PCA) was added to the dataset. Again, ANNs models were developed and their performance indicators were obtained. Additionally, a t-test was performed in order to determine if the data distributions of MSE and R were the same as the corresponding to the previous step. If the distributions were not the same and the statistical indicators were better, this second variable was kept. Otherwise, it was discarded. In further steps, new variables (or components) were added to the dataset and the process continued until all the variables (or components) were added to the dataset and tested.
- **NMF:** In this case, the difference relied on the fact that there was not an addition of new variables or components in each step. In contrast, the next rank approximation in ascending order was used.

In the second stage, the four best models (one from each feature selection method) were combined in a fuse ANN approach. In this approach, after finding the best model for each method in stage 1, each model was trained to get an estimation of the entire original dataset. Next, these four estimations were used as inputs of the second stage ANN model, and the best possible model was found. Finally, this model was trained and a final stage-2 estimation of EPSA Algeciras dataset was obtained.

## RESULTS AND DISCUSSION

The results of the experimental procedure are presented in this section. In the first stage, ANNs based on backpropagation with early stopping have been used and here we compare their results with different feature selection procedures in order to determine the best model to estimate the hourly concentration values of $NO_2$ atmospheric pollutant at EPSA Algeciras monitoring station.

Table 3 shows the results obtained in the first stage. The four feature selection methods produce different results. Therefore, different ANN models can be selected in each case (using a different number of hidden units). Based on the values of the performance indexes, the good performance of ANNs models with early stopping is demonstrated. Table 4 shows the data set used in each best model of the first stage per selection method.

**TABLE 3**
**Results Stage-1. Best ANN models for each feature selection method.**

| Method | nhiddens | R | MSE | RMSE | D | MAE |
|---|---|---|---|---|---|---|
| PCA | 10 | 0.823 | 172.222 | 13.122 | 0.900 | 9.194 |
| NMF | 7 | 0.819 | 175.918 | 13.260 | 0.896 | 9.371 |
| P-values | 11 | 0.834 | 162.639 | 12.751 | 0.904 | 8.898 |
| RELIEFF | 8 | 0.832 | 164.093 | 12.808 | 0.903 | 8.942 |

**TABLE 4**
**Data set used in each ANN model of Table 3.**

| Method | Variables numbers / Components / Rank approximation |
|---|---|
| PCA | 1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 15, 35, 37 (components) |
| NMF | 32 (rank approximation) |
| P-values | 4, 5, 7, 10, 12, 13, 15, 18, 20, 23, 28, 29, 30, 31, 32, 37, 38 (variable numbers) |
| RELIEFF | 3, 5, 6, 7, 8, 10, 12, 15, 16, 18, 19, 20, 21, 23, 28, 29, 30, 31, 32, 33, 36, 37, 38 (variable numbers) |

Table 5 shows the results obtained in the second stage. Stage-2 fuses the results obtained by the ANNs selected in Stage-1 using the best selected ANN models (in Table 2). As we can see, the ANN model of Stage-2 outperforms the models of Stage-1 in all the quality indexes measured. Considering the results, the introduction of the proposed two-stage approach achieved better prediction performance and were shown to be superior to the simple first stage models (composed by a feature selection method + ANN). The proposed procedure guarantees a better prediction performance of the concentration values at the Algeciras monitoring station.

**TABLE 5**
**Results Stage-2. Fusion of the selected models of Stage-1. Best ANN model.**

| nhiddens | R | MSE | RMSE | D | MAE |
|---|---|---|---|---|---|
| 5 | 0.866 | 133.450 | 11.551 | 0.925 | 7.916 |

The results of the estimation of the two-stage approach for EPSA Algeciras monitoring station, using one month of the period as an example, are plotted in Figure 2. A good fit of this approach is shown with a closer adjustment to the observed values. The gaps in the figure indicate missing values

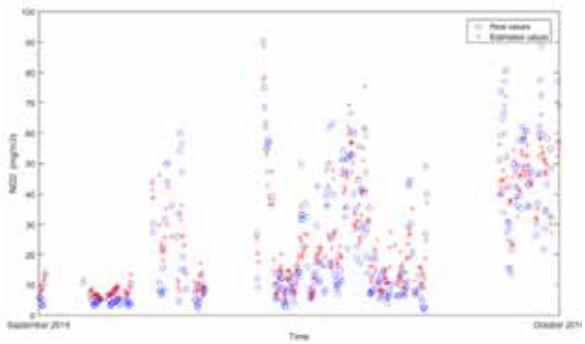

**FIGURE 2**
**Comparison of the observed and estimated values with the best model of the second stage.**

## CONCLUSIONS

A data fusion approach based on ANNs has been presented in this work in order to estimate the hourly concentration values of $NO_2$ at Algeciras (EPSA) monitoring station. A set of models with different inputs (coming from different feature selection methods) and configurations (number of hidden units) have been tested using a resampling procedure in two stages.

The results from the two-stage approach outperform those obtained by the single models. The results obtained demonstrated the utility of this data fusion approach. The application of this approach can become an efficient supporting tool to different purposes such imputation of missing data, automatic detection of decalibration or calculation of concentration values in unseen or unregistered points where no monitoring station is located.

## ACKNOWLEDGEMENTS

This work is part of the coordinated research projects TIN2014-58516-C2-1-R and TIN2014-58516-C2-2-R supported by MICINN (Ministerio de Economía y Competitividad-Spain). Monitoring data has been kindly provided by the Environmental Agency of the Andalucian Government.

## REFERENCES

[1] Akkoyunlu A., Erturk F. (2002) Evaluation of air pollution trends in Istanbul. Int. J. Environ. Pollut. 18(4), 388-398.
[2] EU (2008) 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. Off. J. Eur. Communities 152, 1-43.
[3] Monteiro A., Lopes M., Miranda A.I., Borrego C., Vautard R. (2005) Air pollution forecast in Portugal: a demand from the new air quality framework directive. Int. J. Environ. Pollut. 25(1-4), 4-15.
[4] Gardner M.W., Dorling S.R. (1998) Artificial neural networks (the multilayer perceptron) - A review of applications in the atmospheric sciences. Atmos. Environ. 32(14), 2627-2636.
[5] Nagendra, S.S., Khare M. (2005) Modelling urban air quality using artificial neural network. Clean Technol. Environ. Policy 7(2), 116-126.

[6] Brunelli U., Piazza V., Pignato L., Sorbello F., Vitabile S. (2007) Two-days ahead prediction of daily maximum concentrations of $SO_2$, $O_3$, $PM_{10}$, NO2, CO in the urban area of Palermo, Italy. Atmos. Environ. 41(14), 2967-2995.
[7] Martin M.L., Turias I.J., Gonzalez F.J., Galindo P.L., Trujillo F.J., Puntonet C.G., Gorriz J.M. (2008) Prediction of CO maximum ground level concentrations in the Bay of Algeciras, Spain using artificial neural networks. Chemosphere. 70(7), 1190-1195.
[8] Khaleghi B., Khamis A., Karray F.O., Razavi S.N. (2013) Multisensor data fusion: A review of the state-of-the-art. Inf. Fusion 14(1), 28-44.
[9] Baruque B., Corchado E. (2011) Fusion Methods for Unsupervised Learning Ensembles. Springer, Berlin, Heidelberg.
[10] Turias I.J., González F.J., Martin M.L., Galindo P.L. (2008) Prediction models of CO, SPM and $SO_2$ concentrations in the Campo de Gibraltar Region, Spain: a multiple comparison strategy. Environ. Monit. Assess. 143(1), 131-146.
[11] Muñoz E., Martín M.L., Turias I.J., Jimenez-Come M.J., Trujillo F.J. (2014) Prediction of $PM_{10}$ and $SO_2$ exceedances to control air pollution in the Bay of Algeciras, Spain. Stoch. Environ. Res. Risk Assess. 28(6), 1409-1420.
[12] Guyon I., Elisseeff A. (2003) An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 3,1157-1182.
[13] Bolón-Canedo V., Sánchez-Maroño N., Alonso-Betanzos A. (2013) A review of feature selection methods on synthetic data. Knowl. Inf. Syst. 34(3), 483-519.
[14] Molina L.C., Belanche L., Nebot A. (2002) Feature selection algorithms: a survey and experimental evaluation. In: Proceedings of the 2002 IEEE International Conference on Data Mining. IEEE Comput. Soc., 306-313.
[15] Saeys Y., Inza I., Larranaga P. (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19), 2507-2517.
[16] Robnik-Šikonja M., Kononenko I. (2003) Theoretical and empirical analysis of RelifF and RReliefF. Mach. Learn. 53(1-2), 23-69.
[17] Jolliffe I. (2005) Principal component analysis. In: Everitt B.S., Howell D.C. (Eds.) Encyclopedia of Statistics in Behavioral Science. Wiley Online Library.
[18] Berry M.W., Browne M., Langville A.N., Pauca, V.P., Plemmons R.J. (2007) Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data Anal. 52(1), 155-173.
[19] Rumelhart D.E., Hinton G.E., Williams R.J. (1986) Learning internal representations by error propagation. In: Rumelhart D.E., McClelland J.L. (Eds.) Parallel Distributed Processing - Explorations in the Microstructure of Cognition. MIT Press, Cambridge, 318-362.
[20] Hornik K., Stinchcombe M., White H. (1989) Multilayer feedforward networks are universal approximators. Neural Networks 2(5), 359-366.
[21] Yao Y., Rosasco L., Caponnetto A. (2007) On early stopping in gradient descent learning. Constr. Approx. 26(2), 289-315.
[22] Corchado E., Wozniak M., Abraham A., Carvalho A.C.P.L.F.D., Snásel V. (2014) Recent trends in intelligent data analysis. Neurocomputing. 126, 1-2.
[23] Zhou Z.H., Wu J., Tang W. (2002) Ensembling neural networks: Many could be better than all. Artif. Intell. 137(1-2), 239-263.
[24] Willmott C.J. (1981) On the validation of models. Phys. Geogr. 2(2), 184-194.
[25] Draper N.R., Smith H. (2014) Applied regression analysis. John Wiley & Sons. New York.

**Received: 10.09.2018**
**Accepted: 15.10.2018**

**CORRESPONDING AUTHOR:**

**Javier Gonzalez-Enrique**
University of Cádiz
Department of Computer Science Engineering, Polytechnic School of Engineering, University of Cádiz, Algeciras, Spain.

E-mail: javier.gonzalezenrique@uca.es