Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-018-01644-0

ORIGINAL PAPER

Spatial and meteorological relevance in NO_2 estimations: a case study in the Bay of Algeciras (Spain)

Javier González-Enrique[1] · Ignacio J. Turias[1] · Juan Jesús Ruiz-Aguilar[2] · José Antonio Moscoso-López[2] · Leonardo Franco[3]

Abstract
This study focuses on how to determine the most relevant variables in order to estimate the hourly NO_2 concentrations in a monitoring network located in the Bay of Algeciras (Spain). For each station of the network, artificial neural networks and multiple linear regression have been used to compute hourly estimation models. Meteorological variables and hourly NO_2 concentrations from the nearby stations have been used as inputs, and a feature selection procedure has been applied as a previous step. The different models developed have been statistically compared. The inputs used in the best estimation model for each station were the most important to estimate each hourly NO_2 concentration level. These estimations can be a very useful resource to provide autonomous capacities as automatic decalibration detection or missing data imputation in monitoring networks. Finally, the similarities between stations, according to the relevance of variables, have been analysed with the aid of a hierarchical clustering algorithm.

Keywords Artificial neural networks · Monitoring networks · Air pollution · Feature relevance

1 Introduction

Air pollution is one of the most important problems that affect the quality of living, especially in industrialized and densely populated areas. A poor air quality can produce very negative effects on human health (Tabaku et al. 2011; Gibson 2015; Chiu and Yang 2015), especially on people who belong to susceptible population groups, such as children and elder people (Kolehmainen et al. 2001; European Environment Agency 2013). Therefore, EU and many national environmental agencies have established regulations that limit the concentration levels of atmospheric pollutants with the aim of improving air quality (European Environment Agency 2014).

For the aforementioned reasons, it is necessary to establish a control strategy for air pollutants. Air quality monitoring networks supply information about the actual status of air quality. Environmental monitoring networks can be composed of a variable number of sensors working together. They can perform detailed measurements and provide accurate data of concentrations of airborne pollution, which can be a very valuable resource in order to manage air quality and take corrective actions if needed.

✉ Javier González-Enrique
javier.gonzalezenrique@uca.es

Ignacio J. Turias
ignacio.turias@uca.es

Juan Jesús Ruiz-Aguilar
juanjesus.ruiz@uca.es

José Antonio Moscoso-López
joseantonio.moscoso@uca.es

Leonardo Franco
lfranco@lcc.uma.es

[1] Department of Computer Science Engineering, Polytechnic School of Engineering, University of Cádiz, Avda. Ramón Puyol, s/n, 11202 Algeciras, Cádiz, Spain

[2] Department of Industrial and Civil Engineering, Polytechnic School of Engineering, University of Cádiz, Avda. Ramón Puyol, s/n, 11202 Algeciras, Cádiz, Spain

[3] Department of Computer Science, ETS Computer Science, University of Málaga, Bulevar Louis Pasteur, 35, 29071 Málaga, Spain

Published online: 19 January 2019

According to Bhaskar and Mehta (2010), urban airborne pollution is principally caused by human activities. Its main origins are industrial processes where combustion is present, especially motor vehicles and industries (Bartra et al. 2007).

The main urban air pollutants are CO, NO_x ($NO + NO_2$), hydrocarbons, and particles. Nitrogen dioxide (NO_2) is one of the most relevant air pollutants. Although high-temperature combustion processes produce NO_2 (Rivera et al. 2015), the main quantity of NO_2 in the atmosphere has its origin in the interactions between ozone with nitrogen oxides close to their point of emission, and also with organic radicals along the dispersion path (Finlayson-Pitts and Pitts 2000). This reddish-brown gas, which is very reactive and toxic, plays a major role in atmospheric reactions that produce smog and acid rain in urban areas. Nitrogen dioxide is also considered to be the main single reason for air quality decrease in metropolitan areas (Westmoreland et al. 2007). Hence, control of its concentration levels is one of the most important objectives of environmental agencies.

Air pollutants concentrations may be affected by several factors: local topography, the intensity of emissions, meteorological factors and the distance between receptors and sources of emission (Dominick et al. 2012). Distance from sources of emission is a key factor for pollutant concentrations (Sun et al. 2004). Additionally, meteorological factors may have a very big impact on the quantities of airborne pollutants present in the atmosphere (Banerjee and Srivastava 2011). As Bai et al. (2016) stated, they are essential elements in the process of dispersion of pollutants across the atmosphere and, therefore, they have a great influence in the everyday changes of air pollutants concentrations (He et al. 2013). Finding out the relevance of meteorological variables is a problem of great complexity that depends greatly on local circumstances and the type of pollutant under study (Khedairia and Khadir 2012).

There are different studies where the influence of meteorological factors on pollutants has been discussed. Kourtidis et al. (2002) studied the pollution levels within urban street canyons in Greece and indicated that primary pollutants decreased their concentrations with higher wind speeds due to ventilation (in the case of NO_2, in a smaller amount). Elminir (2005) described in his work the reliance of pollutants on meteorological factors and underlined that wind and relative humidity were the most influencing parameters on airborne pollution. Turias et al. (2008) predicted CO, SPM and SO_2 levels in The Bay of Algeciras (Spain), and pointed out how CO showed a high correlation between wind speed and temperature. In the case of wind direction, their study showed how this correlation was negative. Martín et al. (2008) predicted CO ground levels in the Bay of Algeciras (Spain) and cited temperature, wind direction and wind speed as the most influential meteorological variables over CO concentrations. İçağa and Sabah (2009) analysed the relationship between air pollutants and meteorological factors in Afyon (Turkey) and described how air pollutants were influenced by temperature, inversion (increase in air temperature with elevation) and humidity. However, no dependency between pollutants concentrations and wind velocity or precipitation could be found. In their study, Xu et al. (2011) reported how wind speed and direction had a great influence on pollutants in the North China Plain. Zhang and Batterman (2013) described how low winds and low dispersion situations led to an increase in the concentration of air pollutants. Muñoz et al. (2014) predicted the episodes where PM_{10} and SO_2 levels surpassed legal concentration limits in the Bay of Algeciras and indicated that wind speed and wind direction were the most influencing variables. Zhang et al. (2015) described how wind speed had a negative impact on pollutant concentrations in three megacities of China. In the same line, Xu et al. (2015) pointed out how light winds favoured a raise in pollutant concentrations, while strong winds caused a decrease in them due to dispersion. Finally, Zu et al. (2017) studied the relation of meteorological factors and PM_{10} concentrations and indicated that slow winds, moderate temperature and pressure conditions and the absence of precipitations where associated with episodes of high PM_{10} levels.

There are also other studies where the relationship between meteorological variables and NO_2 concentrations have been specifically studied. Shi and Harrison (1997) studied the possibilities of regression modelling for forecasting NO_2 and NO_x in London using meteorological variables and found that wind speed was an important factor in both NO_x and NO_2 concentrations. Gardner and Dorling (1999) predicted hourly NO_2 and NO_x concentrations in London using meteorological variables. For models with no emission factor, results showed that 47% of the variability of NO_2 and 54% of the variability of NO_x were caused by changes in the meteorological predictors. Kukkonen et al. (2003) undertook an evaluation of artificial neural networks (ANNs) models for forecasting NO_2 and PM_{10} in Helsinki and concluded that climatic factors could have a great effect on the performance of the models obtained. Chen et al. (2009) studied the gaseous pollutants near a traffic line in Beijing and indicated the existence of a negative correlation between wind speed and concentrations of NO_2. Parra et al. (2009) studied the ambient concentrations of NO_2 in northern Spain and highlighted the negative influence of wind speed on pollutant concentrations due to dispersion. Dominick et al. (2012) studied how meteorological components affected concentrations of PM_{10} and NO_2 in Malaysia. Results revealed that wind speed and NO_2 concentrations presented a negative

correlation. Khedairia and Khadir (2012) described how pollutants are greatly influenced by meteorological factors of the study area. In their study focused on Annaba (Algeria), it was found that high NO_2 concentrations were related to lower values of temperature and wind speed.

The main objective of this study is to determine the most relevant meteorological and spatial variables (including the set of monitoring stations) in order to estimate the hourly NO_2 concentrations in a monitoring network located in the Bay of Algeciras (Spain). To achieve this goal, ANNs and MLR were used to develop the estimation models. A set of meteorological variables and NO_2 concentrations were used as inputs and a feature selection procedure was applied as a previous step. Results were statistically analysed in order to find the best model for each station. The variables that were used as inputs of the best estimation model constituted the subset of the original variables that were the most relevant for each case.

The rest of this paper is organized into several sections. Section 2 describes the area of study and the database used. Section 3 gives a brief description of the methods and techniques used in this work. Section 4 presents the experimental design. Section 5 discusses the obtained results. Finally, the conclusions are shown in Sect. 6.

2 Data and area description

The area of study covers the Bay of Algeciras area in the Campo de Gibraltar region, which is located in the southernmost part of the Iberian Peninsula. This area harbours one of the most important industrial zones in Spain, and is densely populated, with almost 300,000 inhabitants in 2018 (Algeciras, 120,000; La Linea, 65,000). It enjoys a Mediterranean climate, and the predominant winds blow from east to west and vice versa. The main industries are an oil-refinery, different petrochemical factories, several power plants and the main stainless-steel factory in Europe. Besides that, the Port of Algeciras is one of the most important ports of the Mediterranean Sea, and there is also an airport located close to the Gibraltar border. Traffic is concentrated in the urban areas of Algeciras and La Línea, but there is also a considerable heavy vehicles traffic related to the import and export operations of the Algeciras Port, where approximately a number of 3.5M TEUs (twenty-foot equivalent unit) are handled and more than 2400 vessels dock per year. It is a very complex scenario, where everything described above is a source of particulate and gaseous air pollution. Despite that, only a few studies have addressed the air pollution problem in the area.

For the present study, the data have been provided by the Environmental Agency of the Andalusian Regional Government, supported by the coordinated research projects TIN2014-58516-C2-1-R and TIN2014-58516-C2-2-R of the Spanish Government. The database covers a period of 6 years (2010–2015) and includes hourly NO_2 concentration values that have been collected by a network of 14 monitoring stations. It also contains meteorological variables, such as relative humidity (%), rainfall (mm/hour), atmospheric pressure (hPa), solar radiation (w/m^2), temperature (C), wind speed (km/h) and wind direction (degrees). These variables have been measured hourly at five meteorological stations. In this study, no methods for missing data imputation have been used.

Figure 1 depicts the location of the study area and the situation of the weather and NO_2 monitoring stations (represented by their codes). The correspondence between stations and their codes is shown in Table 1. Codes one to fourteen indicate NO_2 monitoring stations, whereas codes WE1 to WE5 indicate weather stations.

A number is assigned to each database variable, from number one to number thirty-eight. Variable numbers from 1 to 14 indicate NO_2 concentrations measured in the station of the same code (i.e. variable 1 corresponds to station 1). Variable numbers from 15 to 38 are assigned to different types of meteorological variables. A description of each variable can be found in Table 2.

As was indicated before, all the variables above are expressing hourly values. It is also important to note that, for the sake of concision, in the following sections of this paper the name of each monitoring station will be used to indicate the hourly NO_2 concentrations measured at that particular station.

According to the studies mentioned in the introduction section, winds are among the most relevant factors that affect pollution levels. In the Bay of Algeciras, there are two predominant winds. On the one hand, east winds, including east and east-southeast, which are locally known as “Levante”. On the other hand, west winds, including west-northwest and west-southwest, which are locally known as “Poniente” (Reyes 2015). These winds are moderate, with maximum speeds of 30 km/h and mean speeds of 8 km/h. Additionally, a closer examination reveals how 90% of the time instants show speeds lower than 16 km/h. Situations with east winds tend to last longer compared to those produced west winds, which are much more variable.

3 Methods

The methods and techniques used in this work are briefly explained in this section.

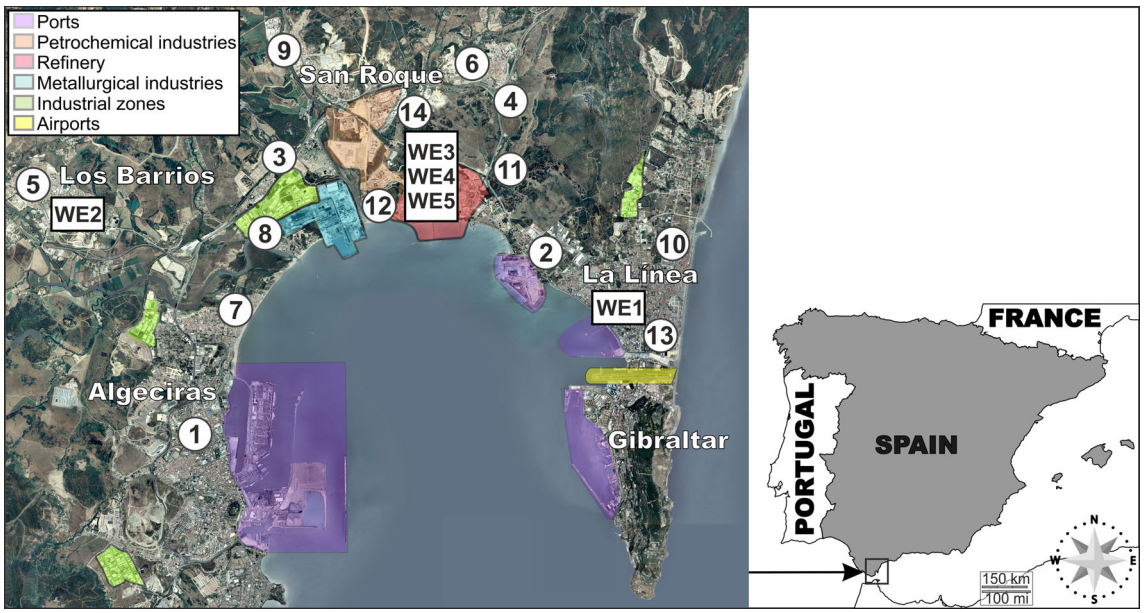

Fig. 1 Location of the Bay of Algeciras (area of study) including the weather and NO_2 monitoring stations

Table 1 NO_2 and weather stations codes

Monitoring/weather station	Code	Monitoring/weather station	Code
EPS Algeciras	1	Economato	11
Campamento	2	Guadarranque	12
Los Cortillijos	3	La Línea	13
Esc. Hostelería	4	Madrevieja	14
Col. Los Barrios	5	La Línea weather station	WE1
Col. Carteya	6	Los Barrios weather station	WE2
El Rinconcillo	7	CEPSA weather station (10 m)	WE3
Palmones	8	CEPSA weather station (60 m)	WE4
Est. San Roque	9	CEPSA weather station (15 m)	WE5
El Zabal	10		

Table 2 Variable description. For meteorological variables, the corresponding weather station is indicated within brackets

Variable	Variable numbers
NO_2 concentration ($\mu g/m^3$)	1–14 (monitoring stations)
Wind direction (°)	15 (WE1), 20 (WE2), 29 (WE4), 32 (WE5)
Wind speed (km/h)	19 (WE1), 31 (WE4), 38 (WE5)
Relative humidity (%)	16 (WE1), 21 (WE2), 25 (WE3), 33 (WE5)
Rainfall (l/m^2)	17 (WE1), 22 (WE2), 26 (WE3), 34 (WE5)
Temperature (°C)	18 (WE1), 30 (WE4), 37 (WE5)
Atmospheric pressure (hPa)	23 (WE2), 27 (WE3), 35 (WE5)
Solar radiation (w/m^2)	24 (WE2), 28 (WE3), 36 (WE5)

3.1 Feature selection

Feature selection approaches can be classified into three different groups: filter, wrapper and embedded methods (Saeys et al. 2007). In filter methods, features are ranked depending on how useful they are for the model, and only the most significant features are kept (Zheng and Zhang 2007). In wrapper methods, the feature selection takes place based on the learning algorithm used to train the model. In this case, different subsets of features are generated and used to train a predictive model. The best subset is selected based on model accuracy (Guyon and Elisseeff

2003). In this study, a combination of filter and wrapper methods have been applied in order to select the best features for each station.

In a first step, variables have been ranked by their ascending regression p values. In a regression analysis, the p value associated with each term is used to test the null hypothesis which states that this coefficient is equal to zero and, consequently, has no effect. An alpha of 0.05 is usually used as the cut-off for significance. If the p value is lower than 0.05, the null hypothesis can be rejected. Thus, there is statistical security that the coefficient chosen by the model is well adjusted.

In a second step, starting with a dataset composed only with the first variable of the ranking, estimation models have been developed. This procedure has been repeated adding new variables to the input data set, according to its position in the ranking, and generating new estimation models with every new addition. Once all the variables have been added, the best model has been selected. Finally, the best subset of variables has been obtained as the input variables used on it.

3.2 Estimation models

Two different methods have been used to develop the estimation models for the NO_2 concentrations: Multiple Linear Regression (MLR) and Feedforward Back Propagation Neural Networks (BPNNs). A brief description of these methods is presented in the next subsections.

3.2.1 Multiple linear regression

MLR is a multivariate statistical method that is used to determine how predictors and the dependent variable are related. It has been widely used in previous pollution forecasting studies (Aguirre-Basurko et al. 2006; İçağa and Sabah 2009; Vlachogianni et al. 2011). The general equation is as follows:

$$y = b_o + \sum_{i=1}^{n} b_i \cdot x_i + \varepsilon \quad (1)$$

where y is the dependent variable, x_i are the independent variables (predictors), b_i are the regression coefficients and ε is the error.

3.2.2 Artificial neural networks

ANNs are computational models inspired by biological neural networks. ANNs have been used in many different areas to inquire about the complex nonlinear relations between predictors and dependent variables. Their use covers predictive modelling, and have been widely used to forecast pollutants in the atmosphere (Chelani et al. 2002; Aguirre-Basurko et al. 2006; Martín et al. 2008; Turias et al. 2008; Muñoz et al. 2014; Russo et al. 2015).

A typical neural network is composed of a number of artificial neurons, called units, which are arranged in different layers and linked by synaptic weights. Feedforward Multilayer Perceptron (MLP) using backpropagation (Rumelhart et al. 1986) is the most popular and widely used design for ANNs. Its architecture is composed of an input layer, one or more hidden layers and an output layer. The network is organized in fully connected layers, with information going forward and errors being propagated backwards in a supervised learning procedure.

Feedforward Neural Networks with a single hidden layer and a sufficiently large number of neurons are able to approximate any nonlinear function and are considered to be universal approximators (Hornik et al. 1989). In this work, BPNN models have been trained using the Levenberg–Marquardt (Marquardt 1963) as the optimization algorithm.

Generalization is a very important issue when using machine learning models, and it can be defined as the ability of the network to offer satisfactory results for unseen new data (Bishop 1995). Hence, reducing the generalization error as much as possible becomes a very important task. In order to avoid overfitting and poor generalization performances, the ANN models were trained using the early stopping technique (Sarle 1995; Gardner and Dorling 1998; Yao et al. 2007). According to this technique, input data is split between a training, a validation, and a test set. It is based on a simultaneous use of a training and validation sets of data while training the network. The validation set is used to evaluate the generalization ability, and this allows the training to be stopped when the maximum generalization capability is reached. The test set is used to evaluate the final performance of the trained ANN.

Finally, it is important to note that the optimal number of neurons in the hidden layer varies depending on the problem that must be solved. Here, a resampling procedure has been used based on a twofold cross-validation. Authors have previously applied this approach successfully in previous works (Turias et al. 2008, 2017; Muñoz et al. 2014).

3.3 Hierarchical clustering

The aim of cluster analysis (Hastie et al. 2009) is to decompose data into different groups where the dissimilarities between the elements within a group are as small as possible and, at the same time, as big as possible between elements belonging to different groups.

Hierarchical clustering (Murtagh 1983; Rokach and Maimon 2005; Hastie et al. 2009) are based on a recursive

partitioning of the elements, which produces a set of nested groups that are arranged as a tree. This process can be performed in two different approaches: agglomerative methods (bottom-up) or divisive methods (top-down).

In this work, an agglomerative hierarchical clustering analysis has been applied in order to find similarities between the relevance of the variables for the monitoring stations. Minimax linkage (Bien and Tibshirani 2011) has been employed to measure dissimilarities between groups. This method supplies prototypical values, which can describe groups and facilitate the interpretation of the analysis. The optimal number of clusters have been decided according to the results obtained by the Calinski–Harabasz's CH index (Calinski and Harabasz 1974), the Davies–Bouldin index (Davies and Bouldin 1979) and the Silhouette method (Rousseeuw 1987).

4 Experimental procedure

The aim of this study is to determine the most relevant variables in order to estimate the hourly NO_2 concentration for each monitoring station. BPNNs and RML where used as estimation methods in three different approaches, depending on the set of input variables used to create the models. In the first one, only the meteorological variables that are present in the initial dataset were used. In the second one, the input dataset was composed exclusively by the NO_2 variables of the stations. Finally, the third approach combined variables of both types in a data-driven scheme (Solomatine et al. 2008). Thus, the whole initial dataset was used to obtain the models (combined approach).

As an initial step, the original data set was pre-processed, and only complete records were used. Therefore, no missing-data imputation procedure was applied to the database, and all the time instants where all variables were not jointly available were removed. As a result, a total of 11,364 time instants of the initial data were used.

For each monitoring station and input variables approach, the input variables were sorted according to their ascending regression p values. Starting with an input dataset composed only of the very best variable, BPNN and RML models were obtained. In every following step, a new variable was added to the input dataset according to its position in the ranking, and its corresponding models were obtained. This process was repeated until all the variables were used as inputs. The BPNNs used included a single hidden layer and were defined using a different number of hidden units (1–20, 25, 30, 35, 40, 45, 50). The early stopping method was used to avoid overfitting and Levenberg–Marquardt was used as the optimization algorithm.

For each model, a random resampling procedure using twofold cross-validation was used as the validation technique. This procedure split the inputs into two non-overlapping folds of equal size (a training and a test set). The parameters of each model were estimated using the training set while the performance was measured using the test set. This process was repeated twice, interchanging training and test sets, and the average of the performance measures was calculated. Due to the random initialization of the weights in ANNs, this validation procedure was repeated 20 times for each model, taking the average of the results. The results of each repetition were also stored in order to perform a multicomparison procedure in a later step. The Pearson correlation coefficient (R) was used in order to evaluate the predictive accuracy of the models. Additionally, the root mean squared error ($RMSE$), the index of agreement (d) and the mean absolute error (MAE) (Willmott 1982) were also calculated. Higher values of d and R indicate a better performance of the models, while lower values of $RMSE$ and MAE imply more precise predictions. These performance indexes are defined in Eqs. (2–5).

$$R = \frac{\sum_{i=1}^{N}(O_i - \bar{O}) \cdot (P_i - \bar{P})}{\sqrt{\sum_{i=1}^{N}(O_i - \bar{O})^2 \cdot \sum_{i=1}^{N}(P_i - \bar{P})^2}} \tag{2}$$

$$RMSE = \left(\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2\right)^{\frac{1}{2}} \tag{3}$$

$$d = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}(|P_i - \bar{O}| + |O_i - \bar{O}|)^2} \tag{4}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - O_i| \tag{5}$$

where P indicates model predicted values and O indicates observed values.

In order to select the best model for each station and input variables approach pair, a multicomparison procedure was performed. Results were analysed using the Friedman test (non-parametric alternative to ANOVA) (Friedman 1937) and the Bonferroni method (Hochberg and Tamhane 1987), in conjunction with the aforementioned performance indexes. The Friedman test evaluated if a significative difference between models was present, while the Bonferroni test determined which of the models were not equivalent. Once the best model is selected for a given monitoring station, the most relevant variables are obtained as the variables used as inputs of the best model.

Finally, with the aim of discovering resemblances in the set of variables that are important for each station, an agglomerative hierarchical clustering algorithm using the Minimax linkage was applied. The optimal number of clusters was determined by the comparison of the results

obtained using the CH index, the Davies–Bouldin index and the Silhouette method.

5 Results and discussion

The results of the experimental procedure are presented in this section. The rankings of variables for each station and proposed input approach, according to their ascending p value, are presented in Fig. 2 (meteorological variables (a) and NO_2 variables (b)) and Fig. 3 (combined approach).

In the case of the meteorological approach, the ranking shows how relative humidity variables and wind direction variables reach the top positions. This prominent group of variables also include wind speed and temperature measured in WE4. Hence, these variables are more likely to be present in the final best models. This result is coincident with many of the works that were revised in the introduction section. In the case of the NO_2 approach, it is clearly observed how the variables that have better positions for a given station are mainly among those variables that are in its own neighbourhood. For the combined approach, the same trend seems to be true. In this case, NO_2 concentrations measured in the neighbouring stations are among the top positions for each station. Nevertheless, meteorological variables such as temperature, relative humidity, wind speed and wind direction are also present in the higher positions of the rankings. Thus, for a certain station, the top variables group is generally composed of a mix of NO_2 and meteorological variables.

Tables 3 and 4 show the results obtained for the proposed approaches using MLR and ANNs respectively and previously unseen data. For each NO_2 monitoring station, different ANN models can be selected varying the number of hidden units (nh) used. For each station and estimation method pair, the best model was selected after applying the Friedman test and the Bonferroni method, as was explained in Sect. 4.

The main goal of this work is to discover the most relevant variables in order to estimate the NO_2 concentrations in each monitoring station. For each case, they were obtained as the inputs used in its best estimation models. A complete list of relevant variables per monitoring station and input approach is shown in Table 5. The knowledge of these variables can be very useful to discover which monitoring stations are related to each other and are affected by changes in the same variables.

Based on this performance indexes of Tables 3 and 4, ANN models with early stopping are shown to be superior to MLR models and give better results in all the cases studied. This can be explained by the ability of the ANN to capture complex, nonlinear relationships among variables. In contrast, MLR models are simpler and only capture linear relations between variables.

The analysis of the best models provided by the meteorological approach using ANNs shows how R values range from 0.60 to 0.77 and d values go from 0.73 to 0.86. Only two stations present d values of 0.85 or more and R values greater than 0.75 (stations 1 and 13). Hence, the overall performance of this approach can only be classified

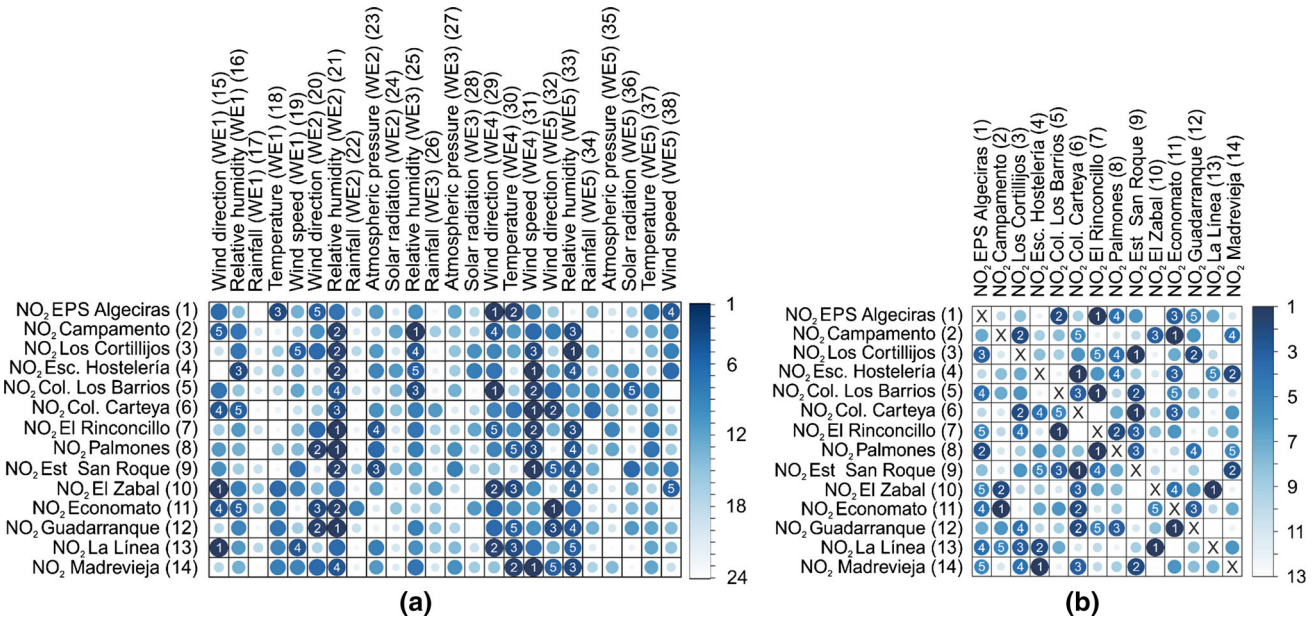

Fig. 2 Ranking per monitoring station based on regression p values using **a** meteorological variables as inputs **b** NO_2 concentrations as inputs. A value equal to 1 means the best variable (lower p value) and 24 (in **a**) or 13 (in **b**) indicate the worst variable (higher p value). The top 5 variables are indicated for each case

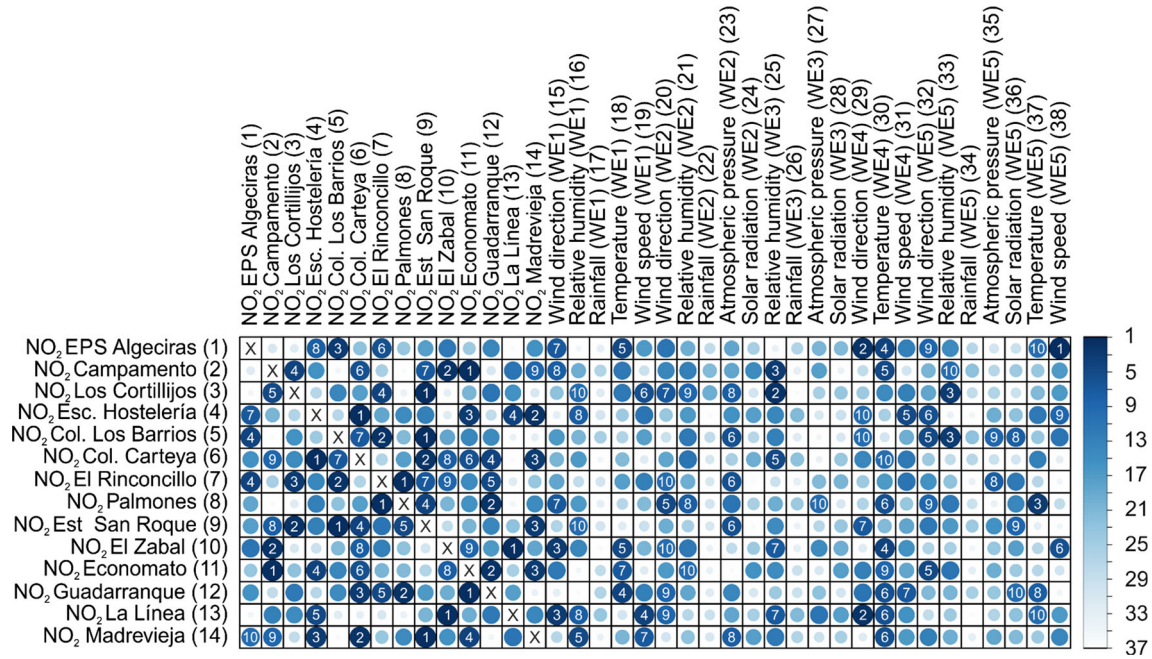

Fig. 3 Ranking per monitoring station based on regression p values using the combined approach for the input dataset (1 means the best variable—lower p value, and 37 means the worst variable—higher p value). The top 10 variables are indicated for each case

as modest. A closer look to Table 4 shows how none of the best models uses less than 9 input variables. This indicates that the top variables of the meteorological rankings are used in all the cases (see Fig. 2a and Table 5). Therefore, relative humidity and wind direction, with the addition of wind speed measured in WE4, constitute the most relevant meteorological features for this approach.

In contrast, the NO_2 input approach using ANNs achieves much better results, with R values between 0.74 and 0.88, and d values between 0.83 and 0.93. The analysis of the input variables used in the best models (see Table 5) shows how mainly neighbouring stations are included, as was expected according to the rankings. However, station 1, which is located in Algeciras, is present in all the models. This station corresponds to the most important city in the area, with NO_2 emissions mainly from urban traffic, heavy vehicles and exhaust gases from ships. Additionally, stations 3, 6 and 11, all of them located very close to the petrochemical industries, are present in at least 10 of the best models.

If both approaches are compared, it is clear that a scheme based on NO_2 inputs outperforms the meteorological approach. It produces models with higher R and d values and lower error indices (*RMSE* and *MAE*). The difference is very appreciable in all the stations except for two specific cases (stations 1 and 3) where results are fairly close. Therefore, only the very top models using the meteorological approach produce results that are comparable to those obtained using the NO_2 approach.

In the case of the combined approach, both types of variables have been ranked and used as inputs of the estimation models (see Fig. 3). Based on the performance indices shown in Table 4, the good estimation performance of the combined models is demonstrated, with R values between 0.78 and 0.89 and d values ranging from 0.87 to 0.94. Compared to the previous approaches, it produces the best results for all the indices and stations. The improvement is remarkable when compared to the meteorological approach, and noticeable in most of the stations if it is compared to the NO_2 approach. In this last case, although it performs better, the difference in stations 6, 7, 9 is not as appreciable. This suggests that much of the information provided by the meteorological variables is already present in the NO_2 variables used.

Scatter plots and the correlation coefficients of estimated versus observed values for stations 1 and 13, for a period of 4 months (July 2014–October 2014), are presented in Fig. 4. The estimations were obtained using the best models for each station (ANNs and the combined approach). These monitoring stations have been selected due to their close location to the most populous cities in the area of study (see Fig. 1).

The estimation model (a) shows an R-value on 0.841, which leads to a determination coefficient (R^2) of 0.707. In the case of estimation model (b), it shows an R-value of 0.877 and a corresponding R^2 of 0.769. The coefficient of determination expresses the proportion of the variance in the dependent variable that is predictable from the

Table 3 Best MLR models for each NO_2 monitoring station (number of variables per input approach: NO_2 ST = 13, meteorological MET = 24, combined ST + MET = 37)

Station	Inputs	Selected variables	$\bar{R}$	$\overline{RMSE}$	$\bar{d}$	$\overline{MAE}$
1	ST + MET	25	0.807	13.657	0.882	9.992
	MET	16	0.714	16.183	0.815	12.171
	ST	13	0.693	16.662	0.798	12.694
2	ST + MET	23	0.759	9.803	0.850	6.573
	MET	16	0.561	12.458	0.674	8.758
	ST	13	0.732	10.257	0.828	6.993
3	ST + MET	19	0.794	9.004	0.876	6.220
	MET	26	0.619	11.654	0.731	7.825
	ST	12	0.739	9.995	0.835	6.748
4	ST + MET	23	0.835	9.580	0.904	6.672
	MET	20	0.554	14.511	0.671	10.340
	ST	10	0.823	9.902	0.896	6.857
5	ST + MET	25	0.800	7.702	0.880	5.223
	MET	18	0.511	11.028	0.623	7.486
	ST	9	0.776	8.099	0.863	5.479
6	ST + MET	18	0.879	6.891	0.932	4.503
	MET	20	0.543	12.108	0.655	8.156
	ST	10	0.874	6.996	0.930	4.576
7	ST + MET	24	0.880	8.286	0.933	5.572
	MET	18	0.625	13.619	0.739	9.919
	ST	8	0.873	8.519	0.929	5.661
8	ST + MET	19	0.860	9.436	0.920	5.999
	MET	22	0.652	14.009	0.762	9.964
	ST	10	0.838	10.091	0.906	6.341
9	ST + MET	25	0.879	6.425	0.932	4.137
	MET	21	0.554	11.220	0.667	7.327
	ST	10	0.870	6.644	0.927	4.282
10	ST + MET	20	0.853	8.515	0.916	6.048
	MET	12	0.652	12.368	0.763	9.027
	ST	6	0.839	8.868	0.907	6.321
11	ST + MET	22	0.838	7.696	0.906	5.242
	MET	23	0.614	11.117	0.729	7.650
	ST	11	0.797	8.507	0.878	5.837
12	ST + MET	22	0.778	11.325	0.867	8.083
	MET	16	0.509	15.523	0.633	11.418
	ST	11	0.753	11.865	0.847	8.399
13	ST + MET	24	0.858	10.825	0.919	7.804
	MET	19	0.708	14.896	0.811	11.170
	ST	6	0.83	11.773	0.901	8.537
14	ST + MET	21	0.848	6.407	0.913	4.306
	MET	20	0.563	9.997	0.676	7.026
	ST	7	0.829	6.758	0.900	4.545

independent variables. The more variance accounted by the models, the closer the data points will be to the perfect fit line. In consequence, there will always exist a difference, or bias, between estimated and the real values according to the R^2 values that the models can achieve. The exception to that statement would be the theoretical case of $R^2 = 1$, a perfect match that represents a 45° degrees line. Hence, a no-perfect match, as occurs here, shows a bias vs the best perfect linear fit.

The main consequence of this bias is that our estimations will reflect only 70.7% and 76.9% of the variance of the real values. With the available dataset, this is the top

Table 4 Best ANN models for each NO_2 monitoring station (number of variables per input approach: NO_2 ST = 13, meteorological MET = 24, combined ST + MET = 37)

Station	Inputs	Selected variables	nh	$\bar{R}$	$\overline{RMSE}$	$\bar{d}$	$\overline{MAE}$
1	ST + MET	19	7	0.834	12.756	0.905	8.910
	MET	13	12	0.754	15.198	0.850	11.010
	ST	9	9	0.765	14.902	0.858	10.580
2	ST + MET	21	9	0.785	9.335	0.874	6.110
	MET	14	13	0.646	11.512	0.765	7.757
	ST	11	9	0.740	10.138	0.837	6.859
3	ST + MET	15	18	0.837	8.145	0.905	5.269
	MET	12	16	0.730	10.176	0.833	6.446
	ST	12	6	0.758	9.688	0.851	6.377
4	ST + MET	23	9	0.848	9.262	0.914	6.377
	MET	15	17	0.638	13.452	0.760	9.250
	ST	10	7	0.832	9.683	0.903	6.695
5	ST + MET	20	10	0.821	7.363	0.895	4.857
	MET	18	14	0.606	10.243	0.731	6.867
	ST	9	7	0.786	7.951	0.872	5.211
6	ST + MET	15	12	0.895	6.443	0.943	3.986
	MET	20	15	0.678	10.638	0.795	6.738
	ST	11	6	0.888	6.642	0.938	4.118
7	ST + MET	18	7	0.893	7.870	0.940	5.236
	MET	15	17	0.704	12.429	0.813	8.756
	ST	6	9	0.885	8.135	0.937	5.365
8	ST + MET	12	9	0.873	9.013	0.929	5.748
	MET	15	16	0.716	12.926	0.824	8.949
	ST	9	13	0.842	9.971	0.910	6.283
9	ST + MET	18	11	0.887	6.234	0.938	3.818
	MET	15	11	0.654	10.226	0.771	6.394
	ST	10	4	0.877	6.488	0.931	3.999
10	ST + MET	16	14	0.875	7.914	0.931	5.482
	MET	12	17	0.721	11.325	0.826	7.942
	ST	5	8	0.851	8.565	0.915	6.033
11	ST + MET	19	16	0.860	7.208	0.921	4.693
	MET	12	17	0.707	9.998	0.817	6.563
	ST	12	8	0.821	8.054	0.896	5.344
12	ST + MET	22	13	0.824	10.225	0.899	7.110
	MET	14	15	0.701	13.011	0.816	9.215
	ST	13	14	0.775	11.428	0.865	8.020
13	ST + MET	17	20	0.877	10.155	0.931	7.156
	MET	9	25	0.771	13.444	0.862	9.556
	ST	6	9	0.840	11.448	0.908	8.092
14	ST + MET	21	8	0.860	6.171	0.921	4.113
	MET	19	19	0.661	9.108	0.783	6.284
	ST	7	6	0.838	6.596	0.907	4.436

performance that the applied models can achieve. A possible increase in the performance (and thus, a lower bias between observed vs. predicted values) could be achieved if new variables adding relevant information were added.

A period of 100 h has been selected from the temporal range mentioned before (July 2014–October 2014). A representation of observed vs. estimated values for this period and stations 1 and 13 is depicted in Fig. 5a and b.

Table 5 Most relevant variables for each NO_2 monitoring station

NO_2 monitoring station number	Inputs	Selected variables (in order)
1	ST + MET	38, 29, 5, 30, 18, 7, 15, 4, 32, 37, 20, 10, 31, 12, 33, 14, 19, 9, 23
	MET	29, 30, 18, 38, 20, 15, 21, 37, 31, 25, 23, 27, 19
	ST	7, 5, 11, 8, 12, 9, 13, 3, 6
2	ST + MET	11, 10, 25, 3, 30, 6, 9, 15, 14, 33, 13, 18, 21, 20, 24, 4, 38, 27, 23, 16, 32
	MET	25, 21, 33, 29, 15, 31, 16, 32, 38, 20, 30, 24, 37, 36
	ST	11, 3, 10, 14, 6, 12, 1, 5, 4, 9, 8
3	ST + MET	9, 25, 33, 7, 2, 19, 20, 23, 21, 16, 12, 18, 31, 5, 30
	MET	33, 21, 31, 25, 19, 20, 16, 38, 29, 28, 23, 30
	ST	9, 12, 1, 8, 7, 6, 11, 4, 5, 13, 2, 10
4	ST + MET	6, 14, 11, 13, 31, 32, 1, 16, 38, 29, 19, 37, 9, 25, 8, 23, 21, 24, 7, 35, 2, 26, 36
	MET	31, 21, 16, 33, 25, 38, 29, 28, 23, 36, 24, 32, 35, 19, 18
	ST	6, 14, 11, 8, 13, 7, 3, 2, 9, 1
5	ST + MET	9, 7, 33, 1, 32, 23, 6, 36, 35, 29, 38, 21, 25, 11, 12, 3, 15, 19, 10, 31
	MET	29, 31, 25, 21, 36, 32, 15, 37, 35, 34, 33, 19, 27, 24, 16, 30, 20, 18
	ST	7, 9, 6, 1, 11, 3, 2, 12, 13
6	ST + MET	4, 9, 14, 12, 25, 11, 5, 10, 2, 30, 21, 31, 37, 29, 3
	MET	31, 32, 21, 15, 16, 34, 30, 25, 29, 23, 26, 33, 37, 35, 24, 20, 36, 28, 19, 38
	ST	9, 3, 11, 4, 5, 14, 8, 12, 10, 1, 2
7	ST + MET	8, 5, 3, 1, 12, 23, 9, 35, 10, 20, 31, 36, 19, 4, 11, 32, 30, 18
	MET	21, 31, 33, 23, 29, 20, 25, 37, 35, 30, 16, 28, 32, 19, 38
	ST	5, 8, 9, 3, 1, 11
8	ST + MET	7, 12, 37, 9, 20, 30, 15, 21, 32, 27, 23, 33
	MET	21, 20, 31, 33, 30, 23, 37, 25, 29, 27, 15, 35,16, 19, 18
	ST	7, 1, 9, 12, 14, 4, 5, 6, 3
9	ST + MET	5, 3, 14, 6, 8, 23, 29, 2, 36, 16, 7, 32, 4, 12, 25, 35, 33, 1
	MET	31, 21, 23, 33, 32, 36, 19, 27, 38, 15, 37, 25, 34, 24, 29
	ST	6, 14, 5, 7, 4, 3, 8, 1, 12, 10
10	ST + MET	13, 2, 15, 30, 18, 38, 25, 6, 11, 20, 1, 21, 7, 16, 27, 31
	MET	15, 29, 30, 33, 38, 18, 21, 16, 19, 31, 26, 23
	ST	13, 2, 6, 11, 1
11	ST + MET	2, 12, 14, 4, 32, 6, 18, 10, 30, 21, 33, 15, 29, 5, 25, 24, 20, 28, 31
	MET	32, 21, 20, 15, 16, 29, 33, 18, 22, 30, 28, 31
	ST	2, 6, 12, 1, 10, 5, 4, 14, 8, 13, 7, 3
12	ST + MET	11, 8, 6, 18, 7, 30, 31, 37, 20, 36, 29, 3, 19, 23, 5, 1, 9, 13, 21,10, 33, 32
	MET	21, 20, 32, 33, 30, 29, 18, 25, 16, 31, 36, 37, 34, 38
	ST	11, 6, 8, 3, 7, 2, 1, 10, 5, 4, 9, 14, 13
13	ST + MET	10, 29, 15, 19, 4, 30, 25, 16, 20, 37, 27, 12, 2, 14, 3, 31, 38
	MET	15, 29, 30, 19, 33, 21, 31, 23, 18
	ST	10, 4, 3, 1, 2, 14
14	ST + MET	9, 6, 4, 11, 16, 30, 19, 23, 2, 1, 13, 25, 32, 38, 8, 10, 31, 24, 35, 33, 18
	MET	31, 30, 33, 21, 32, 20, 25, 18, 19, 27, 23, 37, 16, 35, 29, 28, 34, 15, 38
	ST	4, 9, 6, 3, 1, 11, 13

Values were estimated using the best models for each station (ANNs and the combined approach). Examining the figure, an acceptable fit between estimated and observed values is observed, as it is expected according to the R and R^2 values of model (a) and model (b) that were previously discussed.

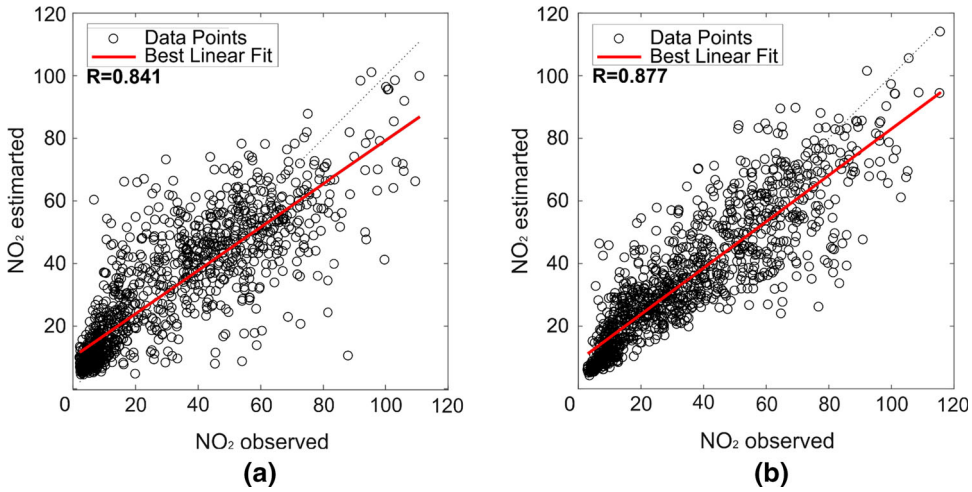

Fig. 4 Scatterplot of observed vs. estimated data and *R* correlation coefficients for EPS Algeciras station (1) (**a**) and La Línea station (13) (**b**) using the best models (July 2014–October 2014). The dashed line corresponds to perfect fit while the red line corresponds to the best linear fit

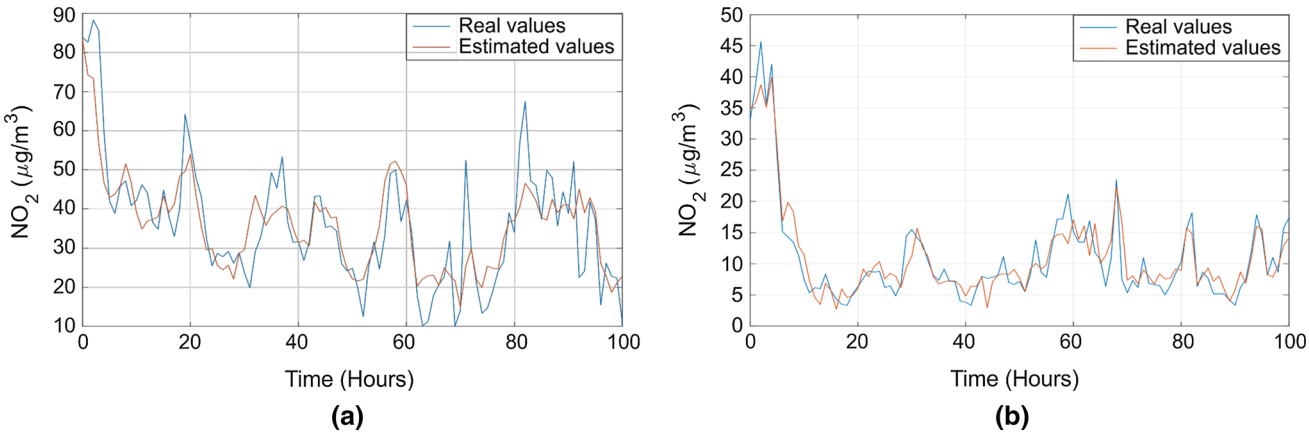

Fig. 5 Observed vs. estimated NO_2 values for EPS Algeciras station (1) (**a**) and La Línea station (13) (**b**) for a selected period of 100 h from the temporal interval (July 2014–October 2014). The estimation values were obtained using the best models for each monitoring station

After obtaining the most relevant variables, and considering their positions in the ranking, a cluster analysis was performed. Through this analysis, it is possible to discover the similarities between stations according to the variables that are relevant to them. To achieve this goal, a hierarchical cluster analysis was applied using the Minimax linkage. The optimal number of clusters was determined using the CH index, the Davies–Bouldin index and Silhouette method and is shown in Table 6.

Table 6 Optimal number of clusters according to the CH index, the Davies–Bouldin index and the Silhouette method

Method	Optimal number of clusters
CH index	2
Davies–Bouldin index	4
Silhouette	2

Results are presented in Fig. 6 and show how the monitoring stations in the area of study can be divided into a western and an eastern main groups (see Fig. 1). Station 12 and station 2 act as prototypes of their respective clusters and, in most cases, stations share similar relevant variables with other stations that are close to them. As it was expected, the obtained clusters are in agreement with the fact that meteorological factors are very similar in nearby stations. Therefore, this result highlights the importance of the spatial relationships between stations.

6 Conclusions

This paper is focused on the determination of the most relevant variables in order to estimate the NO_2 concentration for each station of a monitoring network. This monitoring network is located in El Campo de Gibraltar region (Spain). To achieve this goal, ANNs and MLR were employed as estimation models, using a set of

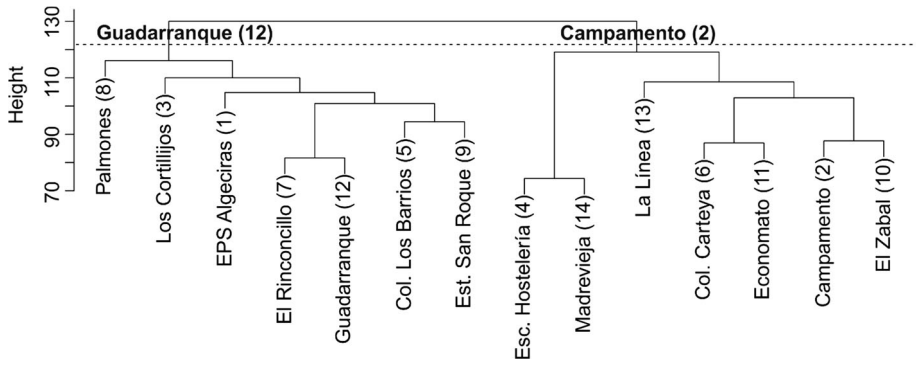

Fig. 6 Cluster dendrogram of the monitoring stations according to the most relevant variables of each monitoring station (ranking positions are considered for each variable). The prototypes are shown in bold

meteorological variables and NO_2 as inputs in three different approaches. The first one was based on the use of meteorological variables only, the second one employed NO_2 variables exclusively and the last one combine variables of both types. After a previous feature selection procedure based on the *p* values of a multiple linear regression as well as an incremental aggregation of variables scheme, the best NO_2 estimation model was obtained for each station and input approach.

Results showed how ANN models using backpropagation outperformed MLR models in every station using different statistical indexes (*R*, *d*, *MSE, MAE*). Focusing on the ANN models, a comparison between the proposed input approaches was performed. Results showed how the combined approach achieved fairly good results and outperformed the other approaches in all the stations and for all the statistical indices. However, it is also important to note that the difference of this approach with respect to the NO_2 approach was limited in a reduced number of cases.

The comparison of the proposed approaches let us understand better the relevance of using meteorological and/or NO_2 variables in the estimation models and drawn some interesting conclusions. As was previously stated, meteorological models are vastly surpassed by the other approaches. They are only close to the NO_2 model in terms of performance if their very best models are considered. However, an improvement in performance can be obtained with the combined approach in all the cases (the amount is important in most cases). This means that relevant information can be added to the models through the use of meteorological variables but the bulk of the important information is contained in the NO_2 variables. Additionally, it is important to note that, in some cases, much of the information provided by the meteorological variables is already present in the NO_2 variables in an implicit way.

The most relevant variables were obtained as the inputs used in the global best estimation model for each station. The knowledge of these groups of relevant variables let us discover which factors affect the NO_2 concentrations at each monitoring station and know which stations are affected by changes in the same variables. A hierarchical cluster was applied in order to discover these similarities between stations. The cluster results revealed the importance of the spatial relationships between stations. This was expected according to the input approaches comparison results and the fact that nearby stations share very similar meteorological conditions. Additionally, based on the list of relevant variables, a map showing the influence area of each of the variables could be obtained.

The application of the estimation of NO_2 measures could produce multiple benefits. It could provide robustness to a network of monitoring sensors. In that sense, in the event of a failure in a monitoring station of the network, its NO_2 measures could be approximated using its corresponding estimation model. Similarly, they might also help in missing values situations producing accurate new values. Furthermore, their use might facilitate the detection of decalibration situations through the comparison of measured data with estimated data provided by the models.

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 have been kindly provided by the Environmental Agency of the Andalusian Government.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Aguirre-Basurko E, Ibarra-Berastegi G, Madariaga I (2006) Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 levels in the Bilbao area. Environ Model Softw 21:430–446

Bai Y, Li Y, Wang X et al (2016) Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. Atmos Pollut Res 7:557–566. https://doi.org/10.1016/j.apr.2016.01.004

Banerjee T, Srivastava RK (2011) Evaluation of environmental impacts of Integrated Industrial Estate—Pantnagar through

application of air and water quality indices. Environ Monit Assess 172:547–560. https://doi.org/10.1007/s10661-010-1353-3
Bartra J, Mullol J, Del Cuvillo A et al (2007) Air pollution and allergens. J Investig Allergol Clin Immunol 17:3–8
Bhaskar BV, Mehta VM (2010) Atmospheric particulate pollutants and their relationship with meteorology in Ahmedabad. Aerosol Air Qual Res 10:301–315. https://doi.org/10.4209/aaqr.2009.10.0069
Bien J, Tibshirani R (2011) Hierarchical clustering with prototypes via minimax linkage. J Am Stat Assoc 106:1075–1084. https://doi.org/10.1198/jasa.2011.tm10183
Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press Inc, New York
Calinski T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat 3:1–27. https://doi.org/10.1080/03610927408827101
Chelani AB, Chalapati RC, Phadke K, Hasan M (2002) Prediction of sulphur dioxide concentration using artificial neural networks. Environ Model Softw 17:161–168. https://doi.org/10.1016/S1364-8152(01)00061-5
Chen J, Wang W, Zhang J et al (2009) Characteristics of gaseous pollutants near a main traffic line in Beijing and its influencing factors. Atmos Res 94:470–480. https://doi.org/10.1016/j.atmosres.2009.07.008
Chiu H-F, Yang C-Y (2015) Air pollution and daily clinic visits for migraine in a subtropical city: Taipei, Taiwan. J Toxicol Environ Health A 78:549–558. https://doi.org/10.1080/15287394.2015.983218
Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell PAMI 1:224–227. https://doi.org/10.1109/tpami.1979.4766909
Dominick D, Latif MT, Juahir H et al (2012) An assessment of influence of meteorological factors on PM10 and NO_2 at selected stations in Malaysia. Sustain Environ Res 22:305–315
Elminir HK (2005) Dependence of urban air pollutants on meteorology. Sci Total Environ 350:225–237. https://doi.org/10.1016/j.scitotenv.2005.01.043
European Environment Agency (2013) Every breath we take: Improving air quality in Europe. Publications Office of the European Union, Luxembourg
European Environment Agency (2014) Annual report 2014 and EMAS environmental statement 2014. Publications Office of the European Union, Luxembourg
Finlayson-Pitts BJ, Pitts JN Jr (2000) Chemistry of the upper and lower atmosphere: theory, experiments, and applications. Academic Press, Cambridge
Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 32:675–701. https://doi.org/10.1080/01621459.1937.10503522
Gardner MW, Dorling SR (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ 32:2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0
Gardner MW, Dorling SR (1999) Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. Atmos Environ 33:709–719. https://doi.org/10.1016/s1352-2310(97)00282-3
Gibson J (2015) Air pollution, climate change, and health. Lancet Oncol 16:e269. https://doi.org/10.1016/S1470-2045(15)70238-X
Guyon I, Elisseeff A (2003) an introduction to variable and feature selection. J Mach Learn Res 3:1157–1182. https://doi.org/10.1016/j.aca.2011.07.027
Hastie TTT, Tibshirani R, Friedman J (2009) The elements of statistical learning, 2nd edn. Springer, New York
He J, Yu Y, Liu N, Zhao S (2013) Numerical model-based relationship between meteorological conditions and air quality and its implication for urban air quality management. Int J Environ Pollut 53:265–286. https://doi.org/10.1504/IJEP.2013.059921
Hochberg Y, Tamhane AC (1987) Multiple comparison procedures. Wiley, New York
Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2:359–366. https://doi.org/10.1016/0893-6080(89)90020-8
İçağa Y, Sabah E (2009) Statistical analysis of air pollutants and meteorological parameters in Afyon, Turkey. Environ Model Assess 14:259–266. https://doi.org/10.1007/s10666-008-9139-5
Khedairia S, Khadir MT (2012) Impact of clustered meteorological parameters on air pollutants concentrations in the region of Annaba, Algeria. Atmos Res 113:89–101. https://doi.org/10.1016/j.atmosres.2012.05.002
Kolehmainen M, Martikainen H, Ruuskanen J (2001) Neural networks and periodic components used in air quality forecasting. Atmos Environ 35:815–825. https://doi.org/10.1016/S1352-2310(00)00385-X
Kourtidis KA, Ziomas I, Zerefos C et al (2002) Benzene, toluene, ozone, NO_2 and SO_2 measurements in an urban street canyon in Thessaloniki, Greece. Atmos Environ 36:5355–5364
Kukkonen J, Partanen L, Karppinen A et al (2003) Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmos Environ 37:4539–4550. https://doi.org/10.1016/S1352-2310(03)00583-1
Marquardt DW (1963) An algorithm for least-squares estimation of nonlinear parameters. J Soc Ind Appl Math 11:431–441. https://doi.org/10.1137/0111030
Martín ML, Turias IJ, González FJ et al (2008) Prediction of CO maximum ground level concentrations in the Bay of Algeciras, Spain using artificial neural networks. Chemosphere 70:1190–1195. https://doi.org/10.1016/j.chemosphere.2007.08.039
Muñoz E, Martín ML, Turias IJ et al (2014) Prediction of PM_{10} and SO_2 exceedances to control air pollution in the Bay of Algeciras, Spain. Stoch Environ Res Risk Assess 28:1409–1420. https://doi.org/10.1007/s00477-013-0827-6
Murtagh F (1983) A survey of recent advances in hierarchical clustering algorithms. Comput J 26:354–359. https://doi.org/10.1093/comjnl/26.4.354
Parra MA, Elustondo D, Bermejo R, Santamaría JM (2009) Ambient air levels of volatile organic compounds (VOC) and nitrogen dioxide (NO_2) in a medium size city in Northern Spain. Sci Total Environ 407:999–1009. https://doi.org/10.1016/j.scitotenv.2008.10.032
Reyes MM (2015) Modelado de alta resolucion para el estudio de la respuesta oceanica al forzamiento del viento en el Estrecho de Gibraltar (Unpublished doctoral dissertation). University of Cádiz, Spain
Rivera C, Stremme W, Barrera H et al (2015) Spatial distribution and transport patterns of NO_2 in the Tijuana–San Diego area. Atmos Pollut Res 6:230–238. https://doi.org/10.5094/APR.2015.027
Rokach L, Maimon O (2005) Clustering methods. In: Maimon O, Rokach L (eds) Data mining and knowledge discovery handbook. Springer, Boston, MA, pp 321–352
Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7
Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, PDP Research Group (eds) Parallel distributed processing: explorations in the microstructure of cognition, vol 1. Foundations. MIT Press, Cambridge, MA, pp 318–362

Russo A, Lind PG, Raischel F et al (2015) Neural network forecast of daily pollution concentration using optimal meteorological data at synoptic and local scales. Atmos Pollut Res 6:540–549. https://doi.org/10.5094/APR.2015.060

Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23:2507–2517. https://doi.org/10.1093/bioinformatics/btm344

Sarle WS (1995) Stopped training and other remedies for overfitting. In: Proceedings of 27th Symposium Interface Computer Science *and Statistics*, pp 352–360

Shi JP, Harrison RM (1997) Regression modelling of hourly NO_x and NO_2 concentrations in urban air in London. Atmos Environ 31:4081–4094. https://doi.org/10.1016/S1352-2310(97)00282-3

Solomatine D, See LM, Abrahart RJ (2008) Data-driven modelling: concepts, approaches and experiences. In: Abrahart RJ, See LM, Solomatine DP (eds) Practical hydroinformatics: computational intelligence and technological developments in water applications. Springer, Berlin, pp 17–30

Sun Y, Zhuang G, Wang Y et al (2004) The air-borne particulate pollution in Beijing—concentration, composition, distribution and sources. Atmos Environ 38:5991–6004. https://doi.org/10.1016/j.atmosenv.2004.07.009

Tabaku A, Bejtja G, Bala S et al (2011) Effects of air pollution on children's pulmonary health. Atmos Environ 45:7540–7545. https://doi.org/10.1016/j.atmosenv.2010.07.033

Turias IJ, González FJ, Martin ML, Galindo PL (2008) Prediction models of CO, SPM and SO_2 concentrations in the Campo de Gibraltar Region, Spain: a multiple comparison strategy. Environ Monit Assess 143:131–146. https://doi.org/10.1007/s10661-007-9963-0

Turias IJ, Jerez JM, Franco L et al (2017) Prediction of carbon monoxide (CO) atmospheric pollution concentrations using meterological variables. WIT Trans Ecol Environ 211:137–145. https://doi.org/10.2495/AIR170141

Vlachogianni A, Kassomenos P, Karppinen A et al (2011) Evaluation of a multiple regression model for the forecasting of the concentrations of NO_x and PM_{10} in Athens and Helsinki. Sci Total Environ 409:1559–1571. https://doi.org/10.1016/j.scitotenv.2010.12.040

Westmoreland EJ, Carslaw N, Carslaw DC et al (2007) Analysis of air quality within a street canyon using statistical and dispersion modelling techniques. Atmos Environ 41:9195–9205. https://doi.org/10.1016/j.atmosenv.2007.07.057

Willmott CJ (1982) Some comments on the evaluation of model performance. Am Meteorol Soc 63:1309–1313. https://doi.org/10.1175/1520-0477(1982)063%3c1309:SCOTEO%3e2.0.CO;2

Xu WY, Zhao CS, Ran L et al (2011) Characteristics of pollutants and their correlation to meteorological conditions at a suburban site in the North China Plain. Atmos Chem Phys 11:4353–4369. https://doi.org/10.5194/acp-11-4353-2011

Xu J, Yan F, Xie Y et al (2015) Impact of meteorological conditions on a nine-day particulate matter pollution event observed in December 2013, Shanghai, China. Particuology 20:69–79. https://doi.org/10.1016/j.partic.2014.09.001

Yao Y, Rosasco L, Caponnetto A (2007) On early stopping in gradient descent learning. Constr Approx 26:289–315. https://doi.org/10.1007/s00365-006-0663-2

Zhang K, Batterman S (2013) Air pollution and health risks due to vehicle traffic. Sci Total Environ 450–451:307–316. https://doi.org/10.1016/j.scitotenv.2013.01.074

Zhang H, Wang Y, Hu J et al (2015) Relationships between meteorological parameters and criteria air pollutants in three megacities in China. Environ Res 140:242–254. https://doi.org/10.1016/j.envres.2015.04.004

Zheng H, Zhang Y (2007) Feature selection for high dimensional data in astronomy. Adv Sp Res 41:1960–1964. https://doi.org/10.1016/j.asr.2007.08.033

Zu Y, Huang L, Hu J et al (2017) Investigation of relationships between meteorological conditions and high PM_{10} pollution in a megacity in the western Yangtze River Delta. Air Qual Atmos Health, China. https://doi.org/10.1007/s11869-017-0472-1