SCIENTIFIC REPORTS
natureresearch

Check for updates

OPEN

Evolution and forecasting of PM10 concentration at the Port of Gijon (Spain)

Fernando Sánchez Lasheras[1]✉, Paulino José García Nieto[1], Esperanza García Gonzalo[1], Laura Bonavera[2] & Francisco Javier de Cos Juez[3]

The name PM_{10} refers to small particles with a diameter of less than 10 microns. The present research analyses different models capable of predicting PM_{10} concentration using the previous values of PM_{10}, SO_2, NO, NO_2, CO and O_3 as input variables. The information for model training uses data from January 2010 to December 2017. The models trained were autoregressive integrated moving average (ARIMA), vector autoregressive moving average (VARMA), multilayer perceptron neural networks (MLP), support vector machines as regressor (SVMR) and multivariate adaptive regression splines. Predictions were performed from 1 to 6 months in advance. The performance of the different models was measured in terms of root mean squared errors (RMSE). For forecasting 1 month ahead, the best results were obtained with the help of a SVMR model of six variables that gave a RMSE of 4.2649, but MLP results were very close, with a RMSE value of 4.3402. In the case of forecasts 6 months in advance, the best results correspond to an MLP model of six variables with a RMSE of 6.0873 followed by a SVMR also with six variables that gave an RMSE result of 6.1010. For forecasts both 1 and 6 months ahead, ARIMA outperformed VARMA models.

The town of Gijón and its Port. Gijón is a town located on the north coast of Spain, in the Principality of Asturias. It is the most populated municipality of this region, with a total of 273,422 inhabitants according to 2016 census. This town, together with Oviedo (220,648 inhabitants) and Avilés (79,514 inhabitants) and other small towns, forms a metropolitan area with more than 850,000 inhabitants. It was founded in the fifth century B.C. During the twentieth century it underwent significant development due to industry, something which is still of great importance to the local economy.

The weather in Gijón is defined by its proximity to the sea and the low mean altitude. The annual level of precipitation is quite high, with a total of 920 L per square meter and year. Regarding temperature, the coldest month is January, with an average temperature of 8.9 °C, while the hottest is August with 19.7 °C. The average annual temperature is 13.8 °C. Winds are sporadic and seasonal. The wind regime is dominated by two main components[1]. During winter it blows from W-WSW, while in summer it comes from E-ENE on the coast.

The Port of Gijón, named *El Musel*, is one of the main ports of the Atlantic Arc and the leading port in the movement of solid bulk in Spain. It is located in the Cantabrian Sea (43°34′N, 5°41′W). Figure 1a shows its position on the North Atlantic Spanish coast and Fig. 1b is an aerial picture of the town, where the location of the port can be observed.

The commercial exploitation of this port started in 1907. In the 1990s there was a development plan that doubled its area and which led to a significant increase in its activity. Its infrastructure is adapted to modern market requirements in terms of drafts, springs and storage areas and a range of services with the best standards of quality. It has 415 hectares of land surface and 7,000 linear meters of dock, structured in areas with the appropriate characteristics to serve each kind of traffic, i.e. specialized terminals for solid bulks, liquids and containers, and multi-purpose facilities for various types of traffic.

In the beginning, the exports were mainly iron ore and coal. Subsequently, the port would expand on its breakwaters and piers, and in the 1940s became the main Spanish port in traffic movement. The industrial activity

[1]Department of Mathematics, Faculty of Sciences, University of Oviedo, c/ Federico García Lorca 18, 33007 Oviedo, Spain. [2]Department of Physics, Faculty of Sciences, University of Oviedo, c/ Federico García Lorca 18, 33007 Oviedo, Spain. [3]Department of Mining Exploitation and Prospecting, University of Oviedo, c/ Independencia 13, 33004 Oviedo, Spain. ✉email: sanchezfernando@uniovi.es

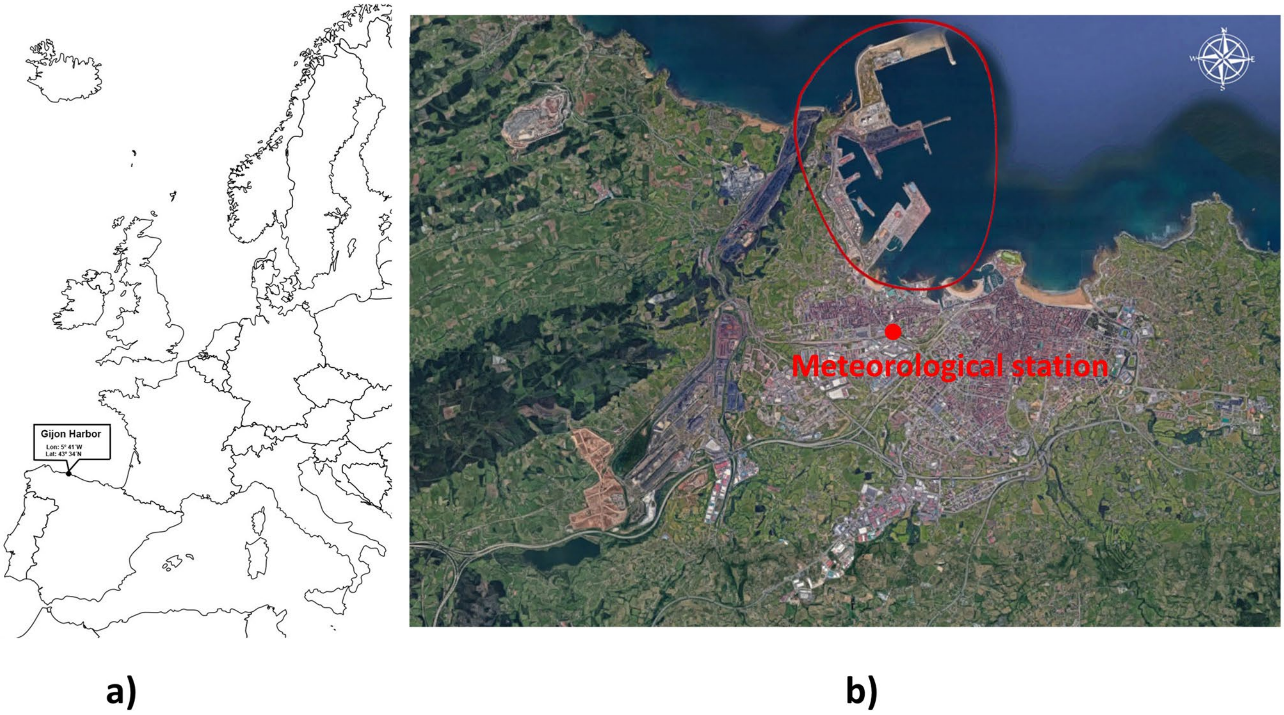

Figure 1. (**a**) position of the Port of Gijon on the North Atlantic coast of Spain, (**b**) aerial picture of Gijón and its Port (inside the red line) including the position of the weather station. *Source*: Google Maps, Map data©2019 Google; https://www.google.es/maps/@43.5547854,-5.6995551,9849m/data=!3m1!1e3. The map was edited with PowerPoint version: 16.0.12527.20260.

of the Principality of Asturias has its main ally in the Port of Gijón. Currently, it is the main bulk port in Spain and one of the most important ports of the Atlantic Arc. According to the traffic statistics of the Annual Report of 2018, a total of 18,226 ships entered the port during that year, which meant a total of 79,294 containers and 12.7 millions of tonnes in the dry bulk terminal, of which 6.4 corresponded to iron ore, 3.4 to iron steel and 2.8 to steam coal. The net revenue in 2018 was 42.2 million euros.

Pollution and particulate matter studies. The World Health Organisation has reported that air pollution has an adverse effect on people's health and development[2]. It is well-known that long-term exposure to high levels of air pollution is linked to decrements in lung function in children[3]. A Swiss study found increased levels of allergic sensitisation in adults living in proximity to busy roads for periods longer than 10 years[4]. Also, the PM_{10} pollutant is amongst those regulated under the Air Quality Framework Directive on ambient air quality assessment and management[5].

A continuous exposure to pollutants such as Carbon Monoxide (CO), Carbon Dioxide (CO_2), oxides of nitrogen (NO_x), and particulate matter is reported to cause health problems in the population living in the affected areas[6,7]. Particulate matter is formed by different chemical products, mostly produced by anthropogenic processes[6] and with significantly variable diameters. Their anthropogenic origin is the reason why they are more present in urban areas[8] than in unpopulated areas.

Air quality issues are relevant in ports and areas nearby. In general, the duty cycle of marine vessels is longer than that of roadside vehicles. This means that ship engines generally use older technology than cars and due to their engine power they are also much more pollutant[9]. Previous studies have analysed PM_{10} concentrations in ports and coastal areas like the Bay of Algeciras in Southern Spain[10]. Another study analysed the impact of $PM_{2.5}$ particles from ship emissions in southern California[11]. In Turkey, shipping emissions in the regions of Candelari Gulf[12] and Ambarli Port[13], both with heavy shipping traffic, were investigated. Research carried out in the port[14] of Tarragona, Spain, made use of multi-linear regression models to study the contribution of different harbour activities to the levels of particulate matter in its area. In the same line there is another study,[15]performed in Barcelona's harbour, also located in Spain and about 80 kms. as the crow flies from Tarragona, which has estimated that around 50–55% of PM_{10} and $PM_{2.5}$ concentrations measured at the port could be attributed to harbour activities and that such activities provide about 9–12% of the total PM_{10} concentration in the air and about 11–15% of $PM_{2.5}$ to the metropolitan area of this city Another interesting and innovative study[16] that deals with the problem of particulate matter in ports was performed in the port of Zhejiang. In this research, with the help of an unmanned aerial vehicle that integrated different sensors, authors have been able to create a profile of the vertical distribution of $PM_{2.5}$, PM_{10} and total suspended particles from ground level to a height of 120 m. A study made at the port of Volos[17], in Greece, found that the highest PM_{10} concentration values were associated with days of calm winds, meaning a wind speed under 0.5 $\frac{m.}{s.}$. The only research into ports that made use of a supervised learning methodology was the one concerning the port of Koper [18]. Koper is the only port

	Minimum	Mean	Maximum	Standard deviation
SO_2 (µg/m^3)	4.0000	7.9706	20.0000	3.2379
NO (µg/m^3)	4.0000	10.9510	30.0000	6.8091
NO_2 (µg/m^3)	7.0000	26.1471	46.0000	8.9159
CO (µg/m^3)	0.1800	0.4023	0.8600	0.1362
O_3 (µg/m^3)	13.0000	38.0000	64.0000	9.9980
PM_{10} (µg/m^3)	18.0000	31.5196	50.0000	7.6271

Table 1. Port of Gijón. Minimum, mean, maximum and standard deviation of the variables of the study: sulfur dioxide (SO_2), nitrogen monoxide (NO), nitrogen dioxide (NO_2), carbon oxide (CO), ozone (O_3) and particulate matter with a diameter less than 10 µm (PM_{10}).

in Slovenia and is located at the northern tip of the Adriatic Sea. Researchers made use of hourly PM_{10} concentrations and employed k-means clustering with Euclidean and city-block distances to cluster days. The results obtained showed the influence of rain intensity and wind speed in the clusters performed but the influence of any other pollutant was not studied. Finally, another study of interest was performed at the Port of Cork, which, like Gijón, is located on the Atlantic coast[19].

Use of machine learning techniques to forecast pollutant concentrations. In general, machine learning can be understood as a subset of methodologies of the artificial intelligence field that are able to learn in an automatic way. In other words, they can learn from data and predict future events. Nowadays, the use of machine learning methodologies has extended to almost all branches of science, including environmental studies. One of the main reasons for the use of machine learning approaches for air quality forecasting is the ability of these methodologies to capture non-linear relationships among variables.

Interest in the forecasting of air pollution in urban area dates back to more than a century ago, when large cities began to have problems with pollution [20]. In the 1970s, several statistical models for pollution forecasting were proposed[21,22]. The first applications of machine learning methodologies in this field were in the 1990s. In those days most research performed made use of artificial neural networks[23,24].

Since then, the different studies performed have made use of other techniques such as genetic algorithms[25], Hierarchical Agglomerative Clustering[26], k-means[27] or support vector machines as regressors[28].

Genetic algorithms have been employed as a supporting methodology for selecting the input variables and designing the high-level architecture of neural networks models. In certain research works[25], they were applied to the selection of the architecture and input variables of a multilayer-perceptron model for forecasting of hourly concentrations of NO. One of the limitations found by this technique is that training each neural network model is a time-consuming task and therefore, the number of parameters to be tuned must be limited.

Hierarchical agglomerative clustering is employed to group objects that are similar in subsets called clusters. The agglomerative clustering methodology starts with many small clusters and merges them together to create large ones. It has been successfully applied in order to study ozone exposure and cardiovascular-related mortality in Canada[26]. The results obtained showed that this methodology is useful for studying the long-term effects of air pollution on cardiovascular diseases.

A recent study has shown how k-means clustering can be employed to categorize different locations in a big and populated city representing the variability of pollution according to the variables employed for the study[27]. Finally, the use of support vector machines as a regressor has also been reported in some studies[28,29]. In one of these[28] the support vector machine is employed as a regressor model for the forecast of the daily Beijing air quality index from 1st January 2014 to mid-2016, while in the others[29,30] they are employed for the forecast of the daily average PM_{10}.

The aim of the present research is to forecast the air quality in a port area, specifically in the port area of the city of Gijón. For this purpose, the article applied different machine learning models (multilayer perceptron neural networks, support vector machine as regressor and MARS) and compared the performance of the predictions obtained for different time intervals with those given by two time series methodologies, one of them univariate (ARIMA) and the other multivariate (VARMA). This means that an exhaustive comparison is made of the prediction from 1 to 6 months in advance of the performance of five methods. This provides an interesting framework for the comparison of methodologies. All these methods were employed in the past for pollution forecasting, but never all in the same research, as far as the authors know. Therefore, the relevance of the present research is that it deals with the topic of monitoring air quality in a city, comparing different machine learning methodologies applied to the same data set.

The database. The information employed for this research has been obtained from one of the meteorological stations belonging to the network of Air Quality Monitoring of the Government of the Principality of Asturias, and more specifically from the one closest to the Port of Gijón, which is located at Argentina Avenue. This station records environmental measurements hourly. As is normal in all this kind of databases, about 0.23% of the raw observations taken each 15 min for all variables were missing. They were imputed with the help of the Multivariate Imputation by Chained Equiations (MICE) algorithm[31].

Table 1 shows the minimum, mean, maximum and standard deviation of the pollutants measured at Gijón Port for the period of study. The values considered for the present research were average monthly measurements from January 2010 to June 2018. Information from January 2010 to December 2017 was employed to forecast values from January to June 2018. Pollutants measured at the Port of Gijón were SO_2, NO, NO_2, CO, O_3 and PM_{10}.

Materials and methods

The present research calculates predictive models of PM_{10} concentration by means of autoregressive integrated moving average (ARIMA), vector autoregressive moving-average (VARMA), multilayer perceptron neural networks (MLP), support vector machines as regressor (SVMR) and multivariate adaptive regression splines (MARS) models. In all cases the PM_{10} values were calculated in two ways: firstly, using the concentration of the six pollutants available as input variables and afterwards employing only four: SO_2, NO, NO_2 and PM_{10}. The main reason why new models using only four variables of the six available are also trained and validated is that many meteorological stations, including some pertaining to the net of Air Quality Monitoring of the Government of the Principality of Asturias are only able to measure these four variables. In other words, the use of only the aforementioned four variables will allow us to compare the model performance according to the input variables employed and will serve as a reference for future studies. Please note that what was said before relates to all the models of the present research except for ARIMA, where only concentration of PM_{10} are employed for the forecasting. In all cases, for continuous variables minimum, mean, maximum and standard deviation were calculated.

Forecasts are performed from 1 to 6 months in advance. The reason why it might be of interest to perform forecasts 6 months in advance is two-fold. On the one hand, high PM_{10} concentrations have adverse effects on human health and on the other, having such a forecast would be helpful in order to take measurements that would make it possible to comply with European air quality standards. According to the results obtained, the best forecast of PM_{10} concentration 1 month ahead is obtained by the SVMR model calculated with six variables. In the case of the forecast 6 months ahead the results of the MLP with six variables are slightly better. In other words, in the short-term the best forecasts are given by SVMR but in the long-term it is outperformed by MLP.

Autoregressive integrated moving average (ARIMA).

ARIMA models can be considered as being an extension of ARMA (autoregressive moving average) known for their ability to provide a parsimonious description of a stationary stochastic process[32]. ARMA models are composed of two polynomial terms, one for autoregression (AR) and another for moving average (MA). Given a time series of data X_t, the ARMA model can be expressed as:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \sigma_i \varepsilon_{t-i}$$

where c is a constant, ε_t are white noise error terms, $\sum_{i=1}^{P} \varphi_i X_{t-i}$ is the autoregressive addend where φ_i are parameters and X_{t-i} is the value of variable X in time $t - i$. $\sum_{i=1}^{qq} \sigma_i \varepsilon_{t-i}$ is the moving-average addend where σ_i are the parameters of the model.

ARIMA models are appropriate for those observation sets that are not necessarily generated by a time series, as is the case of the present problem. They considerably improve the empirical description of non-stationary time series[29]. A stochastic process can be characterized as an ARIMA model if the d-th difference of X_t, constitutes an ARMA stationary and invertible process of p, q orders.

In this case, p represents the order of the autoregressive part of the model, q is the order of the weighted moving average and another parameter called d represents the number of differencing required to reach stationarity[33]. If the differencing operator is denoted by ∇, the general ARIMA equation can be written as follows[30]:

$$\emptyset_p(B)\nabla^d(X_t - L) = \theta_q(B)_{\varepsilon_i}$$

where $\emptyset_p(B)$ and $\theta_q(B)$ are the autoregressive polynomials of weighted moving averages and ε_i is the model perturbation.

$$\emptyset_p(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p$$
$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - B_q B^q$$

A more in-depth explanation of ARIMA models goes beyond the scope of this research and can be found elsewhere[34]. All the models employed in the present research were calculated with the help of the statistical software R[35]. ARIMA models were calculated with the help of the series library[36].

Vector autoregressive moving-average (VARMA).

The Vector autoregression Moving-Average (VARMA) method models the next step in each time series using an ARMA model. In other words, it can be considered the generalization of ARMA to multivariate time series. This kind of model makes it possible to compute a set of time series at the same time, obtaining their within-correlations and cross-correlations[32]. For these models calculus was performed with the help of the MTS library[37].

If a k-dimensional time series is represented by z_t, the vector autoregressive moving-average VARMA $\left(p, q\right)$ process can be expressed as:

$$\phi(B)z_t = \phi_0 + \theta(B)a_t$$

where ϕ_0 is a constant vector

$$\phi(B) = I_k - \sum_{t=1}^{p} \phi_t B_t$$

$$\theta(B) = I_k - \sum_{t=1}^{q} \theta_t B_t$$

are two matrix polynomials and a_t is a sequence of independent and identically-distributed random vectors with mean zero and positive-definitive covariance matrix $\sum_a$.

A general VARMA (p, q) model is represented as follows[37]:

$$z_t = \phi_0 + \sum_{t=1}^{p} \phi_i z_{t-1} + a_t - \sum_{t=1}^{q} \theta_i a_{t-i}$$

In this equation p and q are nonnegative integers, ϕ_0 is a vector of constants, ϕ_i and θ_j are two constant matrix and $\{a_t\}$ is a sequence of independent and identically-distributed random vectors with mean zero and positive definite covariance matrix.

According to Tsay and Wood[37], the VARMA model expressed in the previous equation can be rewritten in a more convenient way as follows:

$$z_t = \phi_0 + \sum_{t=1}^{p} \phi_i z_{t-1} + L b_t - \sum_{t=1}^{q} \theta_j^* b_{t-j}$$

where $\theta_j^* = \theta_j L$ where L is a lower triangular matrix with 1 being the diagonal elements. The determination of p and q values was performed following a methodology suggested in previous research[38]. Akaike information criterion[39] (AIC) and Schwarz information criterion[40] (SIC) were employed to balance the improvement in the value of the log-likelihood function with the loss of degrees of the freedom which results from increasing the lag order of a time series model. With the help of both the maximum $p\ and\ q$ values were calculated. All those models with $p\ and\ q$ values less or equal to then were calculated and finally, those with the best RMSE were presented in this paper.

Multilayer perceptron neural networks (MLP). One of the first bio-inspired machine learning models was the one-layer perceptron. This kind of network was proposed by Rosemblatt[41] as a possible modelization of the neuron of the human brain. The rule of the perceptron adaption consists of a supervised iterative method that modifies the neuron weights. The multilayer perceptron is useful as a way in which to modelize a function. In a neural network the outcome is modelled by an intermediary data set of unobservable variables called hidden variables, which are linear combinations of the original predictors. However, this linear combination is typically transformed by a nonlinear function.

Kolmogorov[42] demonstrated that a two-layer network (one hidden layer and one output layer), with a non-linear differentiable activation function is able to approach any "soft" mapping if the number of neurons in the hidden layer is high enough. If a two-layer network like the one employed in the present research is considered, the operations for a system with p input variables, one output variable and q neurons in the hidden layer can be expressed as:

$$y(n) = \sigma\left(w^y \cdot \varphi\left(w^h \cdot x(n)\right)\right)$$

where $y(n)$ and $x(n)$ are the output and input of the net; σ is the activation function of the output layer; φ is the activation function of the hidden layer; w^y and w^h are the weights matrix for the output and hidden layer respectively.

One main requirement in order to make possible the MLP training[43] is that σ and φ be continuously-differentiable functions. Training is performed with the backpropagation method, which is a recursive application of the gradient descent method. For the purposes of this research, the neural network models were trained and validated with the help of the library neuralnet[44]. The activation function employed is the logistic function. A more in-depth explanation of the foundations of neural networks may be found elsewhere[45].

Support vector machines as regressor (SVMR). Support Vector Machines were introduced by the work of Vapnik[46]. Although they were created by binary classification, nowadays they are used for different kinds of problems. Those employed for regression problems are called SVMR[29].

Let a training data set $S = \left\{(x_1, x_2), \ldots \left(x_n, y_n\right)\right\}$, where $x_i \in \Re^d$ and $y_i \in \Re$ the regression task involves finding those parameters $w = (w_1, \ldots, w_d)$ that make it possible to find the following lineal function[27]:

$$f(x) = w_1 x_1 + \cdots + w_d x_d + b$$

As in practice it is not possible to find these parameters with a prediction error equal to zero, a concept called soft margin is employed. For this, variable ξ_i is employed and the equation is written as follows:

$$\min \frac{1}{2} w, w + c \sum_{i=1}^{n} \left(\xi_i^+ + \xi_i^-\right)$$

Please note that $\xi_i^+ > 0$ when the forecast of the model $f(x_i)$ is larger than its real value y_i and $\xi_i^+ < 0$ in other cases.

With the help of the lagrangian function and the Karush–Kuhn–Tucker conditions, the problem can be expressed as follows:

$$f(x) = \sum_{i=1}^{n} \left(\alpha_i^- - \alpha_i^+\right) x, x_0 + b^*$$

where

$$\alpha_i^+ = C - \beta_i^+$$
$$\alpha_i^- = C - \beta_i^-$$
$$b^* = y_i - w^*, x_i \pm \varepsilon$$

In those cases where data cannot be adjusted with the help of a linear function, kernels are employed[47]. Kernels transform data into a new space called characteristics space.

The regressor associated to the lineal function in the new space is as follows:

$$f(x) = \sum_{i=1}^{n} \left(\alpha_i^- - \alpha_i^+\right) K(x, x_i)$$

please note that b^* is not included in the function as it can be included as a constant inside the kernel. The kind of kernel function to be employed depends on the problem to be solved. For example, the radial basis function has been shown to be very effective, but in those cases where the data set comes from a linear regression, the linear kernel function obtains better results[48]. The SVM as regressor models have been implemented with the functionalities of the library e1071[49]. A good explanation of the use of SVM as regressor can be found in the work of Drucker et al.[50].

Multivariate adaptive regression splines (MARS).

MARS is a non-parametric modelling method driven by the following equation[51]:

$$y_t = f(x_t) = \beta_0 + \sum_{i=1}^{k} \beta_i \cdot B(x_{it})$$

where y_t is the output variable for each time t and β_i are the model parameters for the different x_{it}. β_0 is the intercept and B represents the model basis functions.

One of the main characteristics of the MARS models is that they do not make use of any a priori hypothesis concerning the relationships among the variables[52]. The basis functions are defined as follows:

$$B^- = \begin{cases} (t - x)^q & if\ x < t \\ 0 & otherwise \end{cases}$$
$$B^+ = \begin{cases} (t - x)^q & if\ x \geq t \\ 0 & otherwise \end{cases}$$

q is the power of the basis function as is always a value either equal o larger than zero. In order to adjust a MARS model and decide which basis functions are to be included, MARS makes use of the generalized cross validation (GCV). This represents the root mean squared error divided by a penalty parameter that is defined by the model complexity[53]. Its equation is as follows:

$$C(M) = M + 1 + d \cdot M$$

where M represents the number of basis functions in the equation and d is a penalty parameter for each base function included in the model. For this research, a value of 2 has been assigned to such a parameter, while the maximum number of tracer interaction type base functions is restricted to 3. The MARS models employed in this research are based on those programmed in the library earth[54]. A complete explanation of MARS models can be found in the original work of Friedman[51]. Also, an easy-to-read introduction to this methodology can be found in the works of Put et al.[55].

Results and discussion

Table 2 shows the Pearson's correlation coefficients of all the variables in the study. The largest correlation coefficient in absolute value corresponds to variables NO and NO_2 with 0.8626, followed by NO and O_3 with − 0.7593 (inverse relationship) and SO_2 and NO_2 and SO_2 and NO with 0.7160 and 0.7090 respectively. Correlation coefficients of variables SO_2, NO, NO_2 and O_3 with PM_{10} can be considered in absolute value terms as moderate as they range from 0.4320 (CO and PM_{10}) to 0.5251 (NO_2 and PM_{10}).

	NO	NO_2	CO	O_3	PM_{10}
SO_2	0.7090	0.7160	0.6503	− 0.5483	0.4923
NO		0.8626	0.6587	− 0.7593	0.5068
NO_2			0.6755	− 0.5475	0.5251
CO				− 0.4823	0.4320
O_3					− 0.4663

Table 2. Pearson's correlation coefficients of the variables of the study.

	Jan-18	Feb-18	Mar-18	Apr-18	May-18	Jun-18
	22.2217	31.5194	19.8269	20.1082	37.0095	31.8833
		32.0564	21.4559	22.9140	32.8949	29.8194
			19.7957	23.4279	34.4069	30.4898
				22.9000	33.2961	29.6945
					34.6428	29.3402
						29.6487
Avg	29	27	26	31	29	24

Table 3. Port of Gijón. Results of the ARIMA models using variable PM_{10}.

p	q	Jan-18	Feb-18	Mar-18	Apr-18	May-18	Jun-18
4	2	39.6108	42.0017	21.0807	39.9202	21.7535	40.9830
4	2		43.1236	24.5948	39.9053	22.4729	41.6383
4	2			21.4770	40.4344	23.8317	34.4105
4	2				40.8564	24.0001	35.0403
4	2					21.8562	33.3678
4	2						32.4333
4	1	40.4407	42.6933	21.5210	40.0030	22.7195	41.5106
4	1		43.8059	24.7208	40.2494	23.4349	41.8195
4	1			21.9345	41.2505	23.9012	34.6527
4	1				41.7434	24.9623	35.1930
4	1					21.9919	33.8758
4	1						32.5900
2	1	40.1493	43.0113	21.3623	39.8731	23.0342	41.7924
2	1		43.9369	25.4061	40.0493	23.9055	41.4342
2	1			22.2033	41.6250	24.4226	34.7485
2	1				41.8263	24.8753	34.8986
2	1					21.7138	34.4839
2	1						32.9979
1	2	39.6597	43.3128	20.8822	39.8591	22.3459	41.5172
1	2		44.7840	25.2376	40.2805	24.0309	40.5961
1	2			21.5377	41.7663	24.4224	34.6140
1	2				41.8746	25.0023	34.8847
1	2					21.5321	34.5682
1	2						32.5428
Avg		29	27	26	31	29	24

Table 4. Port of Gijón. Results of the VARMA models using variables SO_2, NO, NO_2 and PM_{10}.

Table 3 shows the results of the ARIMA model using the previous values of PM_{10} as the input variable. Tables 4, 5, 6, 7 and 8 show the results obtained using the different models of four (SO_2, NO, NO_2 and PM_{10}) and six variables (SO_2, NO, NO_2, CO, O_3 and PM_{10}) employed in the present research. In all cases, the results are presented in the same way. The first line represents the forecast performed using information from January 2010 to December 2017 as training values. This forecast is performed for the following 6 months. The second

p	q	Jan-18	Feb-18	Mar-18	Apr-18	May-18	Jun-18
4	2	37.8290	44.0486	25.6528	40.2896	25.9201	39.4256
4	2		42.8897	29.4979	36.6604	29.4432	39.6890
4	2			24.6193	36.6633	30.2776	31.0831
4	2				37.7013	29.0341	33.1441
4	2					27.4516	33.9081
4	2						31.8559
4	1	38.7335	43.5777	25.5632	40.6248	26.2742	40.3105
4	1		43.6040	28.9298	37.4593	28.6250	39.3706
4	1			24.2618	37.2412	30.0657	31.7601
4	1				38.8584	28.0238	33.7027
4	1					27.9027	35.1452
4	1						32.9623
2	1	39.0075	44.2329	24.5744	41.0882	26.1770	40.7826
2	1		44.0753	28.3803	39.1346	29.0462	38.8954
2	1			24.7798	37.6661	28.4491	32.0509
2	1				39.2261	27.1545	34.1648
2	1					26.8998	34.9791
2	1						32.8426
1	2	39.6361	44.3558	25.1907	41.3853	25.3394	41.8425
1	2		43.5959	27.5150	39.5201	27.3642	39.6072
1	2			24.0832	39.2072	28.1761	33.7224
1	2				40.5151	26.5538	34.6911
1	2					27.1467	34.9188
1	2						33.0177
Avg		29	27	26	31	29	24

Table 5. Port of Gijón. Results of the VARMA models using variables SO_2, NO, NO_2, CO, O_3 and PM_{10}.

	Jan-18	Feb-18	Mar-18	Apr-18	May-18	Jun-18
Model with variables SO_2, NO, NO_2 and PM_{10}						
	22.8982	28.3153	26.0853	20.4793	35.1029	30.9771
		30.2043	26.3742	23.2469	30.1220	31.0470
			25.1822	24.2686	32.2976	29.8388
				25.4043	31.2319	27.8622
					33.6755	27.8836
						29.1743
Model with variables SO_2, NO, NO_2 CO, O3 and PM_{10}						
	23.9208	29.9514	21.0074	22.0775	34.2918	31.4352
		30.3128	19.3318	24.6153	30.3982	30.5587
			24.2857	25.6399	31.9302	29.8496
				26.3989	30.9463	29.5339
					33.3901	29.3019
						29.7334
Avg	29	27	26	31	29	24

Table 6. Port of Gijón. Results of the MLP models with variables SO_2, NO, NO_2 and PM_{10} and with variables SO_2, NO, NO_2, CO, O_3 and PM_{10}.

line shows the forecast performed using information from January 2010 to January 2018 and the forecasts from February 2018 (1 month ahead) to June 2018 (5 months ahead) as training values. For all the cases, and in order to make an easy comparison of real values with forecasting, root mean squared errors (RMSE) forecasting values from 1 to 6 months ahead and 1 month ahead for all models are presented in Table 9. In the case of the ARIMA model (Table 3), the one that only makes use of past PM_{10} concentrations in order to predict their future values, the RMSE obtained for forecasts performed 1 month ahead was 6.3163 while the RMSE for forecast performed from 1 to 6 months ahead, the RMSE value was 7.6312. Please note that when we speak about the RMSE obtained

	Jan-18	Feb-18	Mar-18	Apr-18	May-18	Jun-18
Model with variables SO$_2$, NO, NO$_2$ and PM$_{10}$						
	22.5224	29.4214	21.7977	21.7465	35.8175	29.2032
		31.2132	19.4056	23.6688	32.2857	28.5922
			24.9580	25.5375	32.5812	29.4920
				26.0547	34.0507	28.3719
					34.4723	28.9101
						29.5071
Model with variables SO$_2$, NO, NO$_2$ CO, O3 and PM$_{10}$						
	23.6383	30.0879	21.4260	21.5572	34.4579	30.7213
		30.9299	19.7200	24.4384	32.0464	29.6021
			25.1539	25.5781	33.4582	30.0453
				26.6899	32.6893	29.9452
					32.9072	28.2090
						29.5126
Avg	29	27	26	31	29	24

Table 7. Port of Gijón. Results of the SVMR models with variables SO$_2$, NO, NO$_2$ and PM$_{10}$ and with variables SO$_2$, NO, NO$_2$, CO, O$_3$ and PM$_{10}$.

	Jan-18	Feb-18	Mar-18	Apr-18	May-18	Jun-18
Model with variables SO$_2$, NO, NO$_2$ and PM$_{10}$						
	31.7247	26.2609	27.9016	21.0456	21.7012	41.1329
		25.9874	28.0799	22.6011	22.2278	39.4284
			29.0392	24.4142	23.5083	39.5599
				24.4118	23.5069	39.5599
					23.4797	39.5665
						41.8199
Model with variables SO$_2$, NO, NO$_2$ CO, O3 and PM$_{10}$						
	29.3314	25.8768	32.2319	30.7817	27.1559	39.4833
		25.8768	31.3865	29.9750	26.4461	39.4833
			31.4188	30.0142	26.5028	39.4833
				30.7211	27.1826	39.4833
					27.0815	39.4833
						39.4833
Avg	29	27	26	31	29	24

Table 8. Port of Gijón. Results of the MARS models with variables SO$_2$, NO, NO$_2$ and PM$_{10}$ and with variables SO$_2$, NO, NO$_2$, CO, O$_3$ and PM$_{10}$.

for a forecast performed 1 month ahead, we refer to the values that are in the diagonal of the table (in the case of Table 3: 22.2217, 32.0564, 19.7957, 22.9000, 34.6428 and 29.6487) as they are the ones calculated 1 month ahead. Regarding the forecast from 1 to 6 months ahead, we compare real values with the forecast of the first row of the table from January 2018 to June 2018 (in the case of Table 3: 22.2217, 31.5194, 19.8269, 20.1082, 37.0095 and 31.8833). Please note that the real monthly averaged values from January to June 2018 were 29, 27, 26, 31, 29 and 24 respectively. These values are included in Tables 3, 4, 5, 6, 7 and 8 make comparisons more direct.

The RMSE values achieved 1 and up to 6 months ahead for all the models trained in the present research are shown in Table 9. For forecasting 1 month ahead, the best results are obtained for the six variables of SVMR and MLP models, followed by the same models including only four variables. These results give us the idea that all the variables included in the study have a certain relevance in terms of performing an accurate PM$_{10}$ prediction. After the MLP and SVMR models, according to RMSE values the next best in forecasting 1 month ahead is the ARIMA model, the only one that makes exclusive use of past PM$_{10}$ values in order to forecast future concentrations. The ARIMA model is followed by MARS with six and four variables, while VARMA are the models that give the worst performance.

In the case of a forecast of up to 6 months ahead, the best performance according to RMSE value is also achieved by 6 variables MLP and SVMR models followed by the same models using only four variables. A remarkable change when compared with the forecast 1 month ahead is that the MARS model that includes 6 variables performs better than the ARIMA model. Finally, and as also happened with the forecasts 1 month ahead, the worst performance was shown by the VARMA models.

Model and variables number	RMSE	
	One month ahead	Up to 6 months ahead
ARIMA	6.3162	7.6312
VARMA (p = 4 q = 2) 4 variables	10.1021	11.4189
VARMA (p = 4 q = 1) 4 variables	10.5529	11.7214
VARMA (p = 2 q = 1) 4 variables	10.6211	11.7832
VARMA (p = 1 q = 2) 4 variables	10.7767	11.8007
VARMA (p = 4 q = 2) 6 variables	8.5767	10.8202
VARMA (p = 4 q = 1) 6 variables	9.2802	11.0743
VARMA (p = 2 q = 1) 6 variables	9.5173	11.4786
VARMA (p = 1 q = 2) 6 variables	9.7252	11.9347
MLP 4 variables	4.6209	6.2661
MLP 6 variables	4.3402	6.0873
SVMR 4 variables	4.9249	6.1191
SVMR 6 variables	4.2649	6.1010
MARS 4 variables	8.2575	8.7319
MARS 6 variables	6.7605	6.8725

Table 9. RMSE values 1 and up to 6 months ahead of all the models employed in the present study.

From our point of view, a remarkable fact is that the model performance in terms of RMSE in both 1- and 6-month ahead models is not only linked to the number of variables considered in it, but also to the kind of model selected. In other words, it is possible to find a model of only one variable (ARIMA) that performs better than others that include six variables in both 1- and 6-month ahead predictions (VARMA). Finally, the importance of a variable is very easy to assess with the help of a MARS model. The importance order found for the prediction of PM_{10} was as follows: PM_{10} value in the previous moments, followed by the previous measurements of CO, NO, O_3, SO_2 and NO_2.

The main limitation of this study is that although original data is taken each 15 min, forecasts are performed for average monthly values. The reason why average monthly values were forecasted is that the results obtained by the authors when daily or hourly forecasts were performed were not as stable as the average monthly values. This is due to the influence of the port traffic in the pollution area, which does not follow a fixed cycle like urban traffic. Another limitation to be overcome in future studies is that in order to improve the results obtained it would be of interest to introduce some meteorological variables such as temperature, humidity, pressure, sun radiation, rainfall and wind speed and direction.

Conclusions

The results obtained in this research allow us to say that it is possible to predict PM_{10} concentration with the help of the value of this variable and the concentration of other pollutants by means of statistical and machine learning models. Also, another interesting issue is that as had already been found in previous studies,[56] the use of the concentration of other pollutants helps to obtain a more accurate prediction. In fact, the most accurate results were obtained for two kind of machine learning models, SVMR and MLP, when they made use of the values of the six available variables. The results obtained show how regression-based models like SVMR, MARS and MLP outperform univariate and multivariate time series-based models (ARIMA and VARMA). According to the findings of this paper and other previous ones[29], this is because the short-term relationships among pollutants are stronger than temporal relationships of PM_{10} concentration values with itself and with other variables. In other words, although it is possible to find certain seasonal patterns in monthly average pollutant values, the relationship of PM_{10} with the concentration of other pollutants is more important than the seasonal pattern.

Finally, this research affords the reader the opportunity to compare different machine learning and time series methodologies applied to the same data set to establish whether they are useful for PM_{10} concentration forecasting. If the average monthly values of PM_{10} from January to June 2018 are compared with those corresponding to the same months of the previous year, the RMSE result is 6.8557. This means that in forecasts 1 month ahead, MLP and SVM models of four and six variables and MARS of six variables outperform it. When forecasts are performed 6 months ahead MLP models of four and six variables and SVM of six variables outperform it. Although the proposed methodologies do not always outperform the mere use of the average values of PM_{10} concentrations of the same months of the previous year, they are a useful complementary tool for planning and taking decisions in advance.

Received: 10 January 2020; Accepted: 30 June 2020
Published online: 16 July 2020

References

1. González-Marco, D., Pau Sierra, J., Fernández de Ybarra, O. & Sánchez-Arcilla, A. Implications of long waves in harbour management: the Gijon port case study. *Ocean Coast. Manag.* **51**, 180–201 (2018).

2. World Health Organization. Effects of air pollution on children's health and development: A review of the evidence. (2005).
3. Gauderman, W. J. *et al.* The effect of air pollution on lung development from 10 to 18 years of age. *New Engl. J. Med.* **351**(11), 1057–1067 (2004).
4. Wyler, C. *et al.* Exposure to motor vehicle traffic and allergic sensitization. *Epidemiology* **11**(4), 450–456 (2000).
5. European Commission. Council Directive 1996/62/EC of 27 September 1996 on ambient air quality assessment and management. Official Journal of the European Communities, 55–63 (1996).
6. Ganguly, R., Sharma, D. & Kumar, P. Trend analysis of observational PM_{10} concentrations in Shimla city, India. *Sustain. Cities Soc.* **51**, 101719 (2019).
7. Grange, S. K., Salmond, J. A., Trompetter, W. J., Davy, P. K. & Ancelet, T. Effect of atmospheric stability on the impact of domestic wood combustion to air quality of a small urban township in winter. *Atmos. Environ.* **70**, 28–38 (2013).
8. Yadav, R., Sahu, L. K., Jaaffrey, S. N. A. & Beig, G. Temporal variation of particulate matter (PM) and potential sources at an urban station of Udaipur in Western India. *Aerosol. Air Qual. Res.* **14**, 1613–1629 (2014).
9. Mueller, D., Uibel, S., Takemura, M., Klingelhoefer, D. & Groneberg, D. A. Ships, ports and particulate air pollution—an analysis of recent studies. *J. Occup. Med. Toxicol.* **5**, 6–31. https://doi.org/10.1186/1745-6673-6-3 (2011).
10. Pandolfi, M., Gonzalez-Castanedo, Y., Alastuey, A., de la Rosa, J. D., Mantilla, E., de la Campa, A. S., Querol, X., Pey, J., Amato, F. & Moreno, T. Source apportionment of PM(10) and PM(2.5) at multiple sites in the strait of Gibraltar by PMF: impact of shipping emissions. Environ. Sci. Pollut. R. Int. **18**(2), 260–269. doi: 10.1007/s11356–010–0373–4 (2011).
11. Agrawal, H. *et al.* Primary particulate matter from ocean-going engines in the Southern California Air Basin. *Environ. Sci. Technol.* **43**, 5398–5402 (2009).
12. Deniz, C., Kilic, A. & Civkaroglu, G. Estimation of shipping emissions in Candarli Gulf, Turkey. *Environ. Monit. Assess.* **17**(1–4), 219–228. https://doi.org/10.1007/s10661-009-1273-2 (2010).
13. Deniz, C. & Kilic, A. Estimation and assessment of shipping emissions in the region of Ambarli Port, Turkey. *Environ. Prog. Sustain.* **29**(1), 107–115 (2009).
14. Alastuey, A. *et al.* Contribution of harbour activities to levels of particulate matter in a harbour area: Hada Project-Tarragona Spain. *Atmos. Environ.* **41**(30), 6366–6378 (2007).
15. Pérez, N. *et al.* Impact of harbour emissions on ambient PM_{10} and $PM_{2.5}$ in Barcelona (Spain): evidences of secondary aerosol formation within the urban area. *Sci. Total Environ.* **571**, 237–250 (2016).
16. Shen, J. *et al.* Vertical distribution of particulates within the near-surface layer of dry bulk port and influence mechanism: a case study in China. *Sustainability* **11**(24), 1–16 (2019).
17. Manoli, E. *et al.* Polycyclic aromatic hydrocarbons and trace elements bounded to airborne PM10 in the harbor of Volos, Greece: Implications for the impact of harbor activities. *Atmos. Environ.* **167**, 61–72 (2017).
18. Žibert, J. & Pražnikar, J. Cluster analysis of particulate matter (PM_{10}) and black carbon (BC) concentrations. *Atmos. Environ.* **57**, 1–12 (2012).
19. Healy, R. M. *et al.* Characterisation of single particles from in-port ship emissions. *Atmos. Environ.* **43**, 6408–6414. https://doi.org/10.1016/j.atmosenv.2009.07.039 (2009).
20. Meisner Rosen, C. Businessmen against pollution in late nineteenth century Chicago. *Bus. Hist. Rev.* **69**(3), 351–397 (1995).
21. Desalu, A., Gould, L. & Schweppe, F. Dynamic estimation of air pollution. *IEEE Trans. Automat. Contr.* **19**(6), 904–910. https://doi.org/10.1109/TAC.1974.1100742 (1974).
22. Lamb, R. G. & Neiburger, M. An interim version of a generalized urban air pollution model. *Atmos. Environ.* **5**, 239–264 (1971).
23. Roadknight, C. M., Balls, G. R., Mills, G. E. & Palmer-Brown, D. Modeling complex environmental data. *IEEE Trans. Neural Netw.* **8**(4), 852–862. https://doi.org/10.1109/72.595883 (1997).
24. Spellman, G. An application of artificial neural networks to the prediction of surface ozone concentrations in the United Kingdom. *Appl. Geogr.* **19**(2), 123–136 (1999).
25. Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J. & Kolehmainen, M. Evolving the neural network model for forecasting air pollution time series. *Eng. Appl. Artif. Intell.* **17**(2), 159–167. https://doi.org/10.1016/j.engappai.2004.02.002 (2004).
26. Cakmak, S., Hebbern, C., Vanos, J., Crouse, D. L. & Burnett, R. Ozone exposure and cardiovascular-related mortality in the Canadian Census Health and Environment Cohort (CANCHEC) by spatial synoptic classification zone. *Environ. Pollut.* **214**, 589–599. https://doi.org/10.1016/j.envpol.2016.04.067 (2016).
27. Govender, P. & Sivakumar, V. Application of k-means and hierarchical clustering techniques for analysis of air pollution: a review (1980–2019). *Atmos. Pollut. Res.* **11**(1), 40–56 (2020).
28. Liu, B. C., Binaykia, A., Chang, P. C., Tiwari, M. K. & Tsao, C. C. Urban air quality forecasting based on multi-dimensional collaborative support vector regression (SVR): a case study of Beijing–Tianjin–Shijiazhuang. *PLoS ONE* **12**(7), 1–17 (2017).
29. García Nieto, P. J., Sánchez Lasheras, F., García-Gonzalo, E. & de Cos Juez, F. J. Estimation of PM_{10} concentration from air quality data in the vicinity of a major steelworks site in the metropolitan area of Avilés (Northern Spain) using machine learning techniques. *Stoch. Env. Res. Risk A.* **32**(11), 3287–3298 (2018).
30. Riesgo García, M. V., Krzemień, A., del Campo, M., García-Miranda, C. E. & Sánchez Lasheras, F. Rare earth elements price forecasting by means of transgenic time series developed with ARIMA models. *Resour. Policy.* **59**, 95–102 (2018).
31. Van Buuren, S. & Groothuis-Oudshoorn, K. Mice: multivariate imputation by chained equations in R. . *J. Stat. Softw.* **45**, 1–67 (2011).
32. Ruey, S. T. *Multivariate Time Series Analysis with R and Financial Applications* (Wiley, New York, 2014).
33. Ordóñez, C., Sánchez Lasheras, F., Roca-Pardiñas, J. & de Cos Juez, F. J. A hybrid ARIMA–SVM model for the study of the remaining useful life of aircraft engines. *J. Comput. Appl. Math.* **346**, 184–191 (2019).
34. Peter, J. B. & Davis, R. A. *Introduction to Time Series and Forecasting* (Springer, New York, 2002).
35. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing (Vienna, Austria, 2019). https://www.R-project.org/.
36. Trapletti, A, & Hornik, K. tseries: Time Series Analysis and Computational Finance. R package version 0.10-47.
37. Ruey, S.T. & Wood, D. MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models. R package version 1.0. https://CRAN.R-project.org/package=MTS (2018).
38. Martin, V., Hurn, S. & Harris, D. *Econometric Modelling with Time Series. Specification, Estimation and Testing* (Cambridge University Press, Cambridge, 2013).
39. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**, 716–723 (1974).
40. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
41. Rosenblatt, F. *Principles of Neurodynamics* (Spartan Books, Washington, 1962).
42. Kolmogorov, A. N. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Dokl. Akad. Nauk SSSR* **114**(5), 953–956 (1957).
43. García-Nieto, P. J., Martínez Torres, J., de Cos Juez, F. J. & Sánchez Lasheras, F. Using multivariate adaptive regression splines and multilayer perceptron networks to evaluate paper manufactured using Eucalyptus globulus. *Appl. Math. Comput.* **219**(2), 755–763 (2012).
44. Fritsch, S., Guenther, F. & Wright, M.N. neuralnet: Training of Neural Networks. R package version 1.44.2. https://CRAN.R-project.org/package=neuralnet (2019).
45. Haykin, S. *Neural Networks: A Comprehensive Foundation* (Prentice Hall, Upper Saddle River, 1998).

46. Vapnik, V. *The Nature of Statistical Learning Theory* (Springer, Berlin, 2000).
47. Suárez Sánchez, A., Riesgo Fernández, P., Sánchez Lasheras, F., de Cos Juez, F. J. & García Nieto, P. J. Prediction of work-related accidents according to working conditions using support vector machines. *Appl. Math. Comput.* **218**(7), 3539–3552 (2011).
48. Kuhn, M. & Johnson, K. *Applied Predictive Modeling* (Springer, New York, 2013).
49. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-2. https://CRAN.R-project.org/package=e1071 (2019).
50. Drucker, H., Burges, C., Kaufman, L., Smola, A. & Vapnik, V. Support Vector Regression Machines. *Adv. Neural Inf.* **9**, 155–161 (1997).
51. Friedman, J. H. Multivariate adaptive regression splines. *Ann. Stat.* **19**(1), 1–67. https://doi.org/10.1214/aos/1176347963 (1991).
52. Sánchez Lasheras, F., García Nieto, P. J., de Cos Juez, F., Mayo Bayón, R. & González Suárez, V. A hybrid PCA-CART-MARS-based prognostic approach of the remaining useful life for aircraft engines. *Sensors.* **15**(3), 7062–7083 (2015).
53. de Andrés Suárez, J., Lorca Fernández, P. & Sánchez Lasheras, F. Bankruptcy forecasting: a hybrid approach using Fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS). *Expert Syst. Appl.* **38**(3), 1866–1875 (2011).
54. Milborrow, S. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. earth: Multivariate Adaptive Regression Splines. R package version 5.1.1. https://CRAN.R-project.org/package=earth (2019).
55. Put, R., Xu, Q. S., Massart, D. L. & Vander Heyden, Y. Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure–retention relationship studies. *J. Chromatogr. A* **1055**(1–2), 11–19. https://doi.org/10.1016/j.chroma.2004.07.112 (2004).
56. García Nieto, P. J., Sánchez Lasheras, F., García-Gonzalo, E. & de Cos Juez, F. J. PM$_{10}$ concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: a case study. *Sci. Total Environ.* **621**, 753–761 (2018).

Author contributions

F.S.L. conceived the ideas, F.S.L. and P.J.G.N. designed the study and retrieved the information. F.S.L. and F.J.C.J. trained and validated the machine learning models. F.S.L., L.B. and E.G.G. wrote the draft of the manuscript. L.B. revised the manuscript.

Competing interests

Laura Bonavera acknowledges the PGC 2018 project PGC2018-101948-B-I00 (MINECO/FEDER) and PAPI-19-EMERG-11 (UNIOVI). The rest of authors declare no other competing financial interest.

Additional information

Correspondence and requests for materials should be addressed to F.S.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.