Environ Model Assess (2018) 23:229–247
DOI 10.1007/s10666-017-9578-y

Air Quality Modeling Using the PSO-SVM-Based Approach, MLP Neural Network, and M5 Model Tree in the Metropolitan Area of Oviedo (Northern Spain)

P. J. García Nieto[1] · **E. García-Gonzalo**[1] · **A. Bernardo Sánchez**[2] · **A. A. Rodríguez Miranda**[2]

Received: 11 April 2017 / Accepted: 17 August 2017 / Published online: 26 August 2017

Abstract The main aim of this study was to construct several regression models of air quality using techniques based on the statistical learning, in the metropolitan area of Oviedo, in northern Spain. In this research, a hybrid particle swarm optimization-based evolutionary support vector regression is implemented to predict the air quality from the experimental dataset (specifically, nitrogen oxides, carbon monoxide, sulfur dioxide, ozone, and dust) collected from 2013 to 2015 in the metropolitan area of Oviedo. Furthermore, a multilayer perceptron network (MLP) and the M5 model tree were also fitted to the experimental dataset for comparison purposes. Finally, the predicted results show that the hybrid proposed model is more robust than the MLP and M5 model tree prediction methods in terms of statistical estimators and testing performances.

Keywords Support vector machines (SVMs) · Particle swarm optimization (PSO) · Artificial neural networks (ANNs) · M5 model tree · Air quality control and modeling

1 Introduction

Air pollution can be defined as the introduction into the atmosphere of chemicals, particulates, or biological elements that can cause discomfort, disease, and even death to humans, animals, or plants. It can also deteriorate the natural or built environment [1–3]. Air pollution has many different sources: (a) natural sources such as volcanic eruptions and windblown dust; (b) static man-made sources such as factories or power plants, or dry-cleaning and degreasing operations; and (c) mobile man-made sources such as motorized vehicles, planes, and trains, all of which contribute to air pollution. Air pollution can be of natural or human origin.

In air quality control, the first response to a known or potential threat to the established air quality standard or guideline is to reduce it. State Implementation Plans (SIPs) formalize such responses in Spain [1–3]. Air pollution is an important environmental problem in metropolitan areas [1–5] like Oviedo (Principality of Asturias, Spain). It may cause health problems that lead to difficulty in breathing, coughing, and worsening of existing cardiac and respiratory problems [3–5]. For instance, diesel exhaust (DE) is one of the main sources of emission of particulate matter originated during combustion. DE has been linked to an increase in thrombosis and acute vascular dysfunction in several human health studies. This would explain the link between increased cardiovascular morbidity and mortality and the previously described particulate matter air pollution [1–3, 6].

Oviedo is the administrative center of the Principality of Asturias in northern Spain. It has a population of 221,202 and covers a land area of 186.65 km^2. It stands at 232 m above sea level and has a population density of 1185.12 inhabitants per square kilometer. The climate of Oviedo, like in the rest of northwest Spain, is more diverse than in other parts of Spain. Summers are generally warm and humid, with sunshine but also some rain. Winters are cold and very wet. Snow is usually present from October to May in the mountains that surround the city. Both rain and occasional snow are regular features in the winters of Oviedo.

The coal-fired power plant in Soto de Ribera is located 7 km south of the city of Oviedo (Fig. 1). This plant power supplies

✉ P. J. García Nieto
lato@orion.ciencias.uniovi.es

[1] Department of Mathematics, Faculty of Sciences, University of Oviedo, 33007 Oviedo, Spain

[2] Department of Mining Technology, Topography and Structures, University of León, 24071 León, Spain

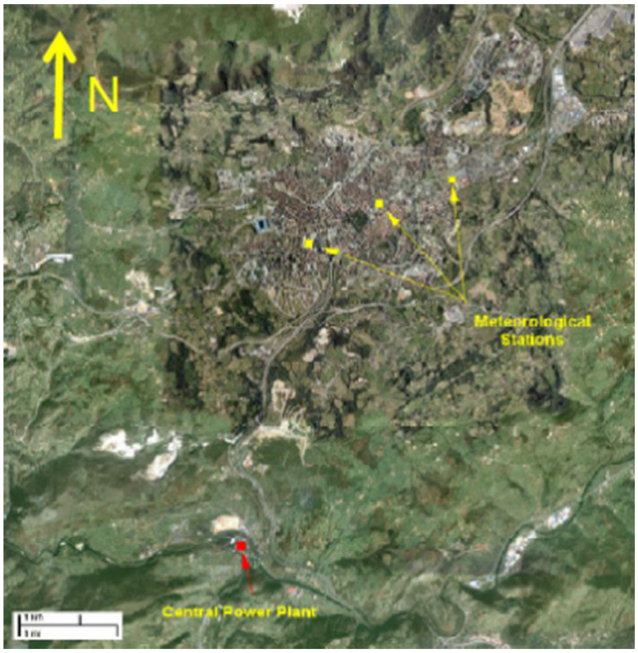

Fig. 1 The geographical location of the meteorological stations in the metropolitan area of Oviedo (northern Spain) and Soto de Ribera power plant (a coal-fired power plant near the city of Oviedo)

most of the electrical energy consumed in Oviedo. The geographical locations of the three meteorological stations and the Soto de Ribera coal-fired power plant are shown in Fig. 1. The Soto de Ribera plant is situated in the district of Ribera de Arriba at an altitude of 126.5 m above sea level.

The monitoring of meteorological pollution, measuring components such as carbon monoxide (CO), sulfur dioxide (SO_2), nitric oxide (NO), nitrogen dioxide (NO_2), ozone (O_3), and particulate matter less than 10 μm (PM_{10}), is becoming increasingly important due to their adverse effects on human health [1–3, 7–11]. Therefore, the EU and many national environmental agencies have established standards and air quality guidelines for permissible levels of these contaminants in the air [5, 11, 12]. The main aim of this work is to build a model for the average daily pollution that would be useful to the authority responsible for air pollution regulation in the corresponding region. The data used for this study has been collected within 3 years, specifically from 2013 to 2015. The numerical experiments applying the PSO-SVM-based technique have obtained good daily modeling accuracy for all pollutants considered. They will be presented and discussed in this paper.

To fix ideas, the aim of this study is to evaluate the application of the support vector machines (SVMs) approach [13–20] in combination with the evolutionary optimization technique known as particle swarm optimization (PSO) [21–24], as well as the multilayer perceptron (MLP) [25–31] and M5 model tree [32–34] to identify the air quality in the metropolitan area of Oviedo (northern Spain) on a local scale, comparing the results obtained. The theoretical support for the learning algorithms of SVMs is given by the statistical learning theory and structural risk minimization. Specifically, five PSO-SVM-based models were created for NO_2, SO_2, and aerosol particles less than 10 μm (PM_{10}) as a function that used the other measured relevant pollutants in air quality as independent variables, namely, NO, CO, and O_3. The purpose was to obtain accurate concentration estimates of the pollutants NO_2, SO_2, and PM_{10} [35–37]. SVM models can be used as an alternative to the classic regression approaches, and they are a new family of models that can be used for estimating values from very different areas [13–20]. The five PSO-SVM-based models were found to improve the accuracy in the case of nonlinear regression problems, such as those related to air quality, which are studied in this paper.

The PSO technique was successfully used here to optimize the tuning of the kernel optimal hyperparameters in the SVM training phase. PSO was introduced by Kennedy and Eberhart in 1995 [21] and is a swarm intelligence (SI) bio-inspired algorithm. The PSO is based on the simulation of the flocking of birds [21–24] and it is similar to other evolutionary computation SI-based algorithms. It also exploits the model of social sharing of information [38, 39]. PSO hybridized with SVM (PSO-SVM) models [38, 39] was used as a learning tool, and trained to estimate the air quality in the metropolitan area of Oviedo from other air pollutants on a local scale.

Model, together with the MLP model and M5 model tree [25–34], was used as automated learning tools, training them in order to predict the air quality in the metropolitan area of Oviedo from the operation physical-chemical input pollutants measured experimentally.

This innovative paper is organized as follows: firstly, the necessary materials and methods to carry out the study are described. Secondly, the results obtained are shown and discussed. Finally, the main conclusions drawn from the results are presented.

2 Materials and Methods

2.1 Sources and Types of Air Pollution

An air pollutant is a substance contained in atmospheric air that can be unhealthy for humans and the environment. Pollutants can be found in the form of solid particles, liquid droplets, or gases. They may be man-made or natural and can be classified as primary or secondary. Mostly, primary pollutants come from a process, such as carbon monoxide from a motor vehicle exhaust, sulfur dioxide from factories, or ash from a volcanic eruption. Secondary pollutants form in the air when primary pollutants interact or react, and therefore, they are not emitted directly. For instance, an important secondary pollutant is ground-level ozone, which is one of the many secondary pollutants which make up photochemical smog [4, 35–37, 40]. Some pollutants can be both primary and secondary, that is, they have been both emitted directly and formed from other primary pollutants.

Human activity produces major primary pollutants such as [1–12, 35–37, 40–42] the following:

- Particulate matter (PM): also called atmospheric particulate matter, or fine particles. These are tiny particles of solids or liquids suspended in a gas. On the other hand, an aerosol would indicate particles and gas together.
- Sulfur oxides (SO_x): in particular, sulfur dioxide, a chemical compound with the formula SO_2. The combustion of coal and petroleum generates sulfur dioxide because these often contain sulfur compounds.
- Nitrogen oxides (NO_x): mainly NO_2 that is emitted during high-temperature combustion. The first product formed is NO, and when NO oxidizes further in the atmosphere, it becomes NO_2.
- Carbon monoxide (CO): is produced by the incomplete combustion of fuels such as coal, wood, or natural gas.

Secondary pollutants include [1–12, 35–37, 40–42] the following:

- Particulate matter: this is composed of gaseous primary pollutants and compounds in photochemical smog. Smog is a special type of air pollution. Typical smog results from large amounts of coal burning in a particular area and is caused by a mixture of smoke and sulfur dioxide.
- Ground-level ozone (O_3): this develops from NO_x and volatile organic compounds (VOCs). Short-term exposure to elevated levels of ozone can be the origin of eye and lung irritations.

Regarding trends in air quality, the Clean Air Act of 1970 established the setting of standards for four of the primary pollutants (aerosols, sulfur dioxide, carbon monoxide, and nitrogen oxides) and the secondary pollutant ozone. Back then, in 1970, these five pollutants were identified as the most widespread and undesirable. Nowadays, lead has been added and they are known collectively as the criteria pollutants and are covered by the United States National Ambient Air Quality Standards (Table 1) [1–12]. The primary standard for each pollutant can be seen in Table 1, which is based on the highest level that can be tolerated by humans without noticeable negative effects, minus a 10–50% margin for safety reasons.

2.2 Experimental Dataset

The government of Asturias, specifically its Section of Industry and Energy, has three air quality monitoring stations located throughout the city of Oviedo (Fig. 1). Every 15 min, measurements are taken of the following primary and secondary pollutants: SO_2, nitrogen oxides (NO and NO_2), CO, PM_{10}, and O_3.

Table 1 National Ambient Air Quality Standards by the United States Environmental Protection Agency (USEPA) [1–12, 40–42]

Pollutant	Maximum allowable concentrations
Carbon monoxide (CO)	
8-h average	9 ppm (10 mg/m^3)
1-h average	35 ppm (40 mg/m^3)
Nitrogen dioxide (NO_2)	
Annual arithmetic mean	0.053 ppm (100 μg/m^3)
Ozone (O_3)	
1-h average	0.12 ppm (235 μg/m^3)
8-h average	0.08 ppm (157 μg/m^3)
Particulate < 10 μm (PM_{10})	
Annual arithmetic mean	50 μg/m^3
24-h average	150 μg/m^3
Sulfur dioxide (SO_2)	
Annual arithmetic mean	0.03 ppm (80 μg/m^3)
24-h average	0.14 ppm (365 μg/m^3)

The six environmental pollutants studied with the aid of these automated monitoring stations were measured with the following sensors: (a) analyzer API 100A for SO_2 gas, (b) analyzer API 200A for NO_x gases, (c) analyzer TELEDYNE 300E for CO gas, (d) analyzer TELEDYNE 400E for O_3 gas, and (e) analyzer DASIBI 7001 for PM_{10} aerosol: this last is based on the reduction of beta rays to measure the concentration of the airborne particulate matter with a diameter less than 10 μm. These sensors collect the data that is processed and delivered on average for the whole city every day. Thus, we have data for the pollutants mentioned above each day, from January 2013 to December 2015. The monthly average concentrations are shown in Table 2.

It is thus possible to study the trend in concentrations of the preceding pollutants in the years 2013, 2014, and 2015 [1–12, 35–37, 40–42].

Figure 2 shows the monthly concentrations of NO_2, SO_2, and CO over 3 years (between 2013 and 2015). The amount of NO_2 fluctuated significantly with several maxima of 51 μg/m^3 in January 2013, 50 μg/m^3 in December 2013, 40 μg/m^3 in January and February 2015, and 46 μg/m^3 in December 2015, respectively. These maxima corresponded to the months of highest energy consumption in homes due to heating and a greater density of cars on the roads during the winter season. Likewise, the minima in the concentration corresponded to the summer months. According to the USEPA Air Quality Standards (Table 1), the maximum permissible concentration of NO_2 expressed as annual arithmetic mean is 100 μg/m^3. The annual arithmetic means for this gas during the years 2013, 2014, and 2015 were 31.8, 27.0, and 34.3 μg/m^3, respectively. Thus, NO_2 concentrations are also below the maximum permitted and meet

Table 2 Monthly average air pollution concentration in the metropolitan area of Oviedo from January 2013 to December 2015

Month of year	SO_2 (μg/m^3)	NO (μg/m^3)	NO_2 (μg/m^3)	CO (mg/m^3)	PM_{10} (μg/m^3)	O_3 (μg/m^3)
January 2013	15	64	51	0.48	31	32
February 2013	14	43	41	0.47	29	36
March 2013	10	30	34	0.40	27	47
April 2013	8	9	22	0.37	33	52
May 2013	9	6	17	0.29	29	54
June 2013	5	8	17	0.35	31	48
July 2013	9	9	19	0.37	36	52
August 2013	13	3	23	0.37	31	48
September 2013	12	20	31	0.38	34	40
October 2013	9	28	33	0.46	28	26
November 2013	13	42	44	0.35	25	22
December 2013	22	71	50	0.43	37	28
January 2014	11	41	40	0.37	27	37
February 2014	8	24	37	0.42	25	46
March 2014	11	25	36	0.42	37	47
April 2014	7	12	29	0.36	25	54
May 2014	7	11	26	0.37	20	59
June 2014	7	9	22	0.36	27	64
July 2014	9	13	18	0.24	24	47
August 2014	7	6	21	0.24	25	35
September 2014	9	17	24	0.26	30	32
October 2014	9	15	21	0.26	28	40
November 2014	7	13	23	0.27	24	32
December 2014	8	32	27	0.32	31	28
January 2015	16	56	40	0.41	29	22
February 2015	9	32	40	0.30	22	32
March 2015	5	17	34	0.31	35	35
April 2015	8	12	31	0.29	39	42
May 2015	5	6	29	0.19	30	42
June 2015	7	31	26	0.22	34	43
July 2015	5	22	31	0.24	33	46
August 2015	7	8	27	0.32	30	41
September 2015	5	11	25	0.47	29	34
October 2015	9	29	39	0.47	28	21
November 2015	15	64	44	0.56	34	26
December 2015	13	68	46	0.62	32	27

air quality standards for a healthy person during these 3 years, including emission peaks.

Similarly, the concentration of SO_2 also fluctuated slightly, with maxima of 20 μg/m^3 in December 2013, 16 μg/m^3 in January 2015, and 15 μg/m^3 in November 2015, respectively. Once more, these maxima corresponded to the winter months. It is also possible to observe that the concentration of SO_2 followed approximately similar behavior to that of the concentration of NO_2, except that the concentration of SO_2 was much smaller. This trend is general throughout the years studied, and it is only logical, as a coal-fired power plant is close to this area (Fig. 1).

Finally, the concentration of CO also went up and down slightly but showed more erratic behavior, and the maxima corresponded to the winter months. Similarly, following the USEPA Air Quality Standards [1–12, 40–42] (Table 1), the maximum permissible concentration of CO expressed as an annual arithmetic mean is 3.33 mg/m^3. The annual arithmetic means for this gas during the years 2013, 2014, and 2015 were 0.39, 0.32, and 0.37 mg/m^3, respectively. Hence, the concentrations of CO during these 3 years, including emission peaks, were below the highest level that can be tolerated by humans, according to USEPA Air Quality Standards [1–12].

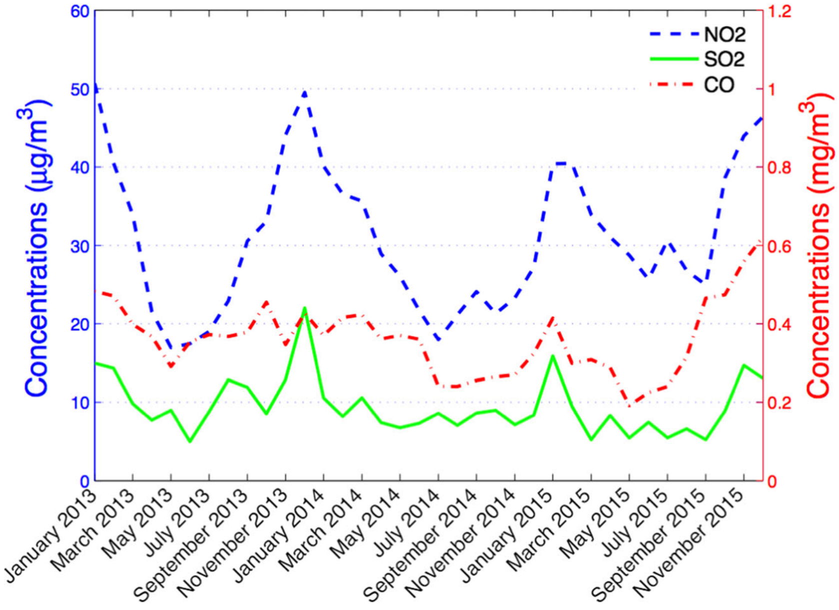

Fig. 2 Monthly trend of nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and carbon monoxide (CO) concentrations during the years 2013, 2014, and 2015 in the metropolitan area of Oviedo

In a similar way, Fig. 3 shows the monthly concentrations of PM_{10}, O_3, and NO from 2013 to 2015 in the metropolitan area of Oviedo. PM_{10} went up and down slightly but remained quite stable at around 30 μg/m^3 with two spikes at 37 μg/m^3 in December 2013 and March 2014, and a minimum of 20 μg/m^3 in May 2014 and a maximum of 39 μg/m^3 in April 2015, respectively. In terms of standard air quality, following the USEPA Air Quality Standards (Table 1), the maximum permissible concentration of PM_{10} expressed as annual arithmetic mean is 50 μg/m^3. The annual arithmetic means for this pollutant during the years 2013, 2014, and 2015 were 30.9, 26.9, and 31.3 μg/m^3, respectively. Therefore, the aerosol concentrations are below the permissible maximum for a healthy person during these 3 years, although emission peaks are close to this value. This behavior can give rise to serious health problems for the population, such as chronic diseases and even death.

Similarly, the concentration of NO fluctuates enormously, with maxima of 64 μg/m^3 in January 2013, 71 μg/m^3 in December 2013 (the highest spike), 56 μg/m^3 in January 2015, and 68 μg/m^3 in December 2015, respectively. Again, these maxima corresponded to the winter months. Furthermore, concentration minima of NO took place during the summer months. Its values were 3 μg/m^3 in August 2013, 6 μg/m^3 in August 2014, and 8 μg/m^3 in August 2015, respectively. Although the initial product of combustion is NO, this gas is rapidly oxidized and converted into NO_2. Its residence time in the atmosphere is

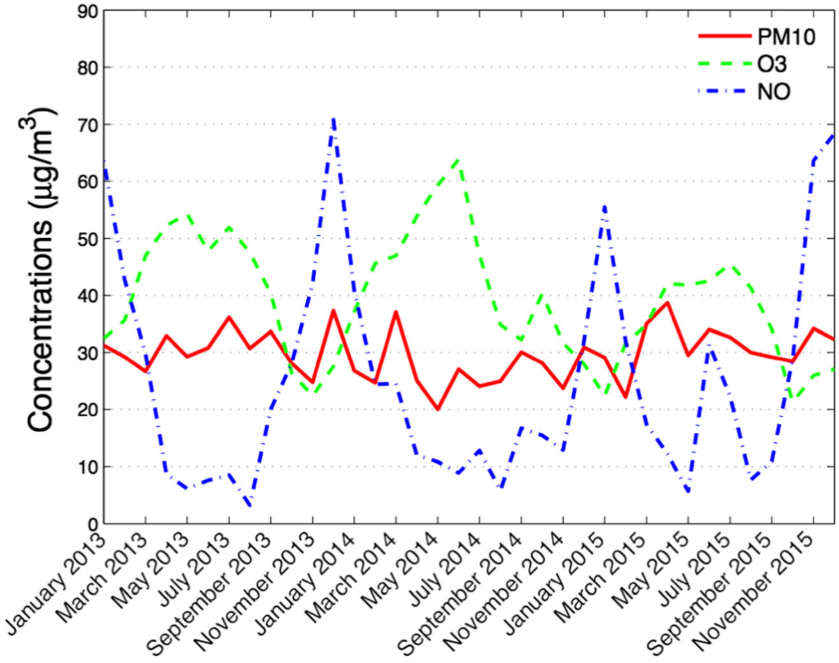

Fig. 3 Monthly trend of particulate matter (PM_{10}), ozone (O_3), and nitric oxide (NO) concentrations during the years 2013, 2014, and 2015 in the metropolitan area of Oviedo

very short and the USEPA Air Quality Standards does not take it into account [1–12].

Finally, the concentration of O_3 also fluctuated considerably, but its behavior is just the opposite of that of NO, that is, maxima of O_3 corresponded to minima of NO and vice versa. This trend is general throughout the years studied, since ozone is associated with photochemical reactions, which require the presence of strong sunlight as a catalyst. The Clean Air Act directs the USEPA to set National Ambient Air Quality Standards for several pollutants, including ground-level ozone, and cities out of compliance with these standards are required to take steps to reduce their levels. In May 2008, the USEPA lowered its ozone standard from 80 to 75 $\mu g/m^3$. This proved controversial, since the agency's own scientists and advisory board had recommended lowering the standard to 60 $\mu g/m^3$, and the World Health Organization recommends 51 $\mu g/m^3$. Many public health and environmental groups also supported the 60-$\mu g/m^3$ standard. The annual arithmetic means for this gas in Oviedo urban area during the years 2013, 2014, and 2015 were 40.4, 43.4, and 34.3 $\mu g/m^3$, respectively. Therefore, the concentrations of this gas were below the maximum permitted, including emission peaks, and meet air quality standards during these 3 years. However, in June 2014, a maximum of 68 $\mu g/m^3$ was reached, therefore exceeding the 60-$\mu g/m^3$ recommendation. This fact could be dangerous for the health of the population of Oviedo. There is a great deal of evidence to show that high concentrations of ozone, created by high concentrations of pollution and daylight UV rays at the Earth's surface, can harm lung function and irritate the respiratory system. Exposure to ozone, and the pollutants that produce it, has been linked to premature death, asthma, bronchitis, heart attack, and other cardiopulmonary problems.

2.3 Support Vector Machine Method

SVMs are a set of supervised learning algorithms closely related to classification and regression problems [13–20]. This last method is called *support vector regression* (SVR). Now, we want to predict a real-valued output $y^{'}$. The regression function $y = f(\mathbf{x})$ for our training data $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^{L}$, where $y_i \in \Re$ and $\mathbf{x}_i \in \Re^D$, with L the number of the samples in the training dataset and D the dimension of the input dataset, is as follows:

$$f(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i + b \tag{1}$$

where $\mathbf{w}$ and b are, respectively, the weight vector and intercept of the model. In general, the regression SVM will use a sophisticated penalty function, not assigning a penalty if the predicted value y_i is less than a distance ε away from the actual value t_i, that is to say, if $|t_i - y_i| < \varepsilon$. The region bound by $y_i \pm \varepsilon$ for all i is called an ε-insensitive tube (Fig. 4). Another modification to the penalty function is that output variables which fall outside the tube are given through two slack variable penalties depending on whether they lie above (ξ^+) or below (ξ^-) the tube (where $\xi^+, \xi^- > 0$ for all i):

$$t_i \le y_i + \varepsilon + \xi^+ \tag{2}$$

$$t_i \ge y_i - \varepsilon - \xi^- \tag{3}$$

The error function for SVR can be written as [13–20]:

$$C \sum_{i=1}^{L} \left(\xi_i^+ + \xi_i^- \right) + \frac{1}{2} \|\mathbf{w}\|^2 \tag{4}$$

where C denotes the *penalty* or *cost parameter* between empirical and generalization errors and ξ_i^+, ξ_i^- are the slack variables defined in Fig. 4. In order to minimize this error function, it is mandatory to take into account the constraints (2) and (3) jointly. To this end, the Karush-Kuhn-Tucker (KKT) optimality conditions [13–20] are applied. These are first-order necessary conditions for a solution in nonlinear programming to be optimal and allowing inequality constraints. If we introduce Lagrange multipliers $\alpha_i^+ \ge 0$, $\alpha_i^- \ge 0$ for all i, the optimization problem for identifying the regression model can be formulated as follows [13–20, 43, 44]:

$$\max_{\alpha^+, \alpha^-} \left[\sum_{i=1}^{L} \left(\alpha_i^+ - \alpha_i^- \right) t_i - \varepsilon \sum_{i=1}^{L} \left(\alpha_i^+ - \alpha_i^- \right) - \frac{1}{2} \sum_{i,j=1}^{L} \left(\alpha_i^+ - \alpha_i^- \right) \left(\alpha_j^+ - \alpha_j^- \right) \mathbf{x}_i \cdot \mathbf{x}_j \right]$$
$$s.t. \left\{ \begin{array}{c} 0 \le \alpha_i^+ \le C \\ 0 \le \alpha_i^- \le C \\ \sum_{i=1}^{L} \left(\alpha_i^+ - \alpha_i^- \right) = 0 \end{array} \right\} \text{for all } i \tag{5}$$

Therefore, new predictions $y^{'}$ can be obtained as:

$$y^{'} = \sum_{i=1}^{L} \left(\alpha_i^+ - \alpha_i^- \right) \mathbf{x}_i \cdot \mathbf{x}^{'} + b \tag{6}$$

In nonlinear cases, we have to proceed by mapping the input low-dimensional vectors via a nonlinear function $\Phi : \mathbb{R}^p \to F$, where F is the feature space of Φ [13–20, 43, 44]. After nonlinear mapping, the regression function has the following form:

$$f(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) + b \tag{7}$$

The solution of this quadratic optimization problem by the Lagrangian dual method [13–20] provides the numerical method with the prediction value:

$$f(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) + b = \sum_{i=1}^{L} \left(\alpha_i^+ - \alpha_i^- \right) K(\mathbf{x}, \mathbf{x}_i) + b \tag{8}$$

where α_i^+, α_i^- are again the Lagrange multipliers of the optimization problem's dual form and $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function satisfying Mercer condition [13–20, 43, 44], and can be described as:

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \Phi(\mathbf{x}_i)^T \Phi\left(\mathbf{x}_j\right) \tag{9}$$

Typical kernel functions described in the bibliography [13–20, 43, 44] are as follows:

- Radial basis function (RBF kernel):

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = e^{-\sigma \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2} \tag{10}$$

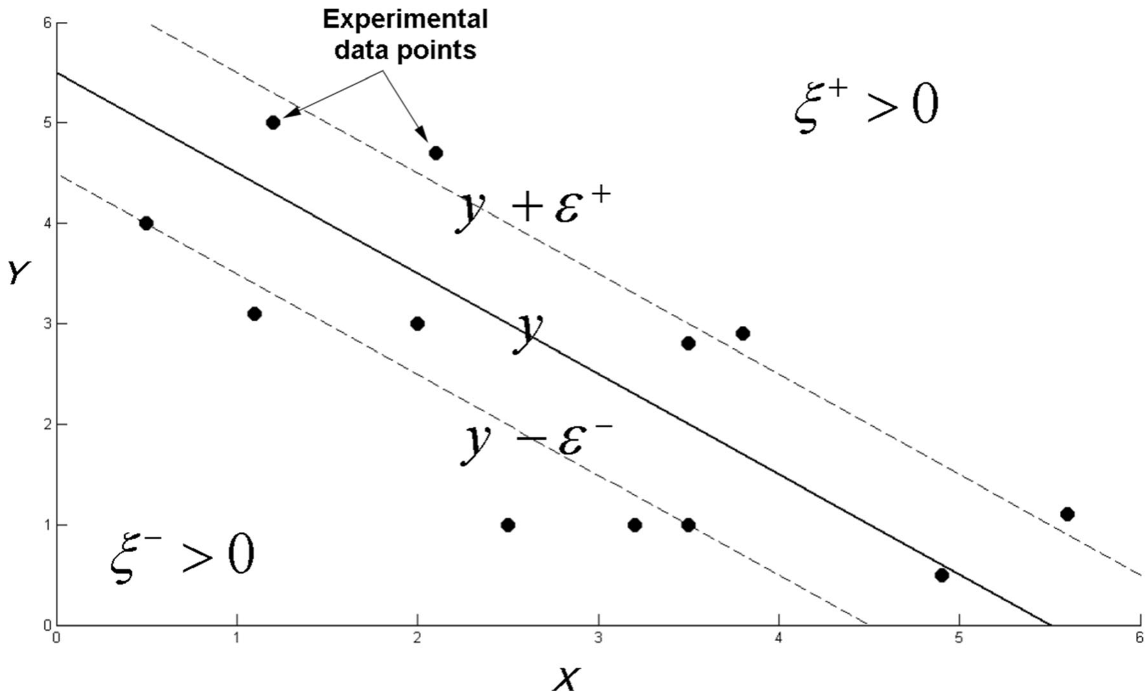

Fig. 4 Regression with ε-insensitive tube for one-dimensional problem

- Polynomial kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\sigma \mathbf{x}_i \cdot \mathbf{x}_j + a)^b \tag{11}$$

- Sigmoid kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\sigma \mathbf{x}_i \cdot \mathbf{x}_j + a) \tag{12}$$

- where a, b, and σ are parameters defining the kernel's behavior.

In summary, to use an SVM to solve a regression problem for data that is not linearly separable, firstly, we need to choose a kernel and relevant parameters that can be expected to map the nonlinearly separable data into a feature space where it is linearly separable.

2.4 The Particle Swarm Optimization Algorithm

PSO is a mathematical optimization/search technique. The PSO is usually used in search spaces with many dimensions. PSO methods were originally attributed to the researchers Kennedy, Eberhart, and Shi [21, 22]. They were initially conceived to elaborate models of social behavior, such as the movement described by living organisms in a flock of birds or a shoal of fish. The algorithm was then simplified and proved to be suitable for solving optimization problems. PSO allows a mathematical problem to be optimized using a population of candidate solutions, denoted as *particles*, moving throughout the search space according to mathematical rules that take into account the position and velocity of the particles. The motion of each particle is influenced by its best local position so far, as well as by the best global positions encountered by other particles as the particles travel through the search space. The theoretical basis of this performance is to make the particle cloud converge quickly to the best solutions. Furthermore, PSO is a metaheuristic technique, as it assumes no hypotheses about the problem to be optimized and can be applied in large spaces of candidate solutions.

Let S be the number of particles in the cloud, each of which has a position $\mathbf{x}_i \in \Re^n$, in the search space and a speed $\mathbf{v}_i \in \Re^n$. Similarly, we will represent the initial position of the particle as $\mathbf{x}_i^0$ and its velocity as $\mathbf{v}_i^0$, both chosen randomly. The best positions correspond to the best values of the fitness function evaluated for each particle. Positions and velocities of each particle are updated taking into account these values, as follows:

$$\mathbf{v}_i^{k+1} = \omega \mathbf{v}_i^k + \phi_1(\mathbf{g}^k - \mathbf{x}_i^k) + \phi_2(\mathbf{l}_i^k - \mathbf{x}_i^k) \tag{13}$$

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k + \mathbf{v}_i^{k+1} \tag{14}$$

The velocity of each particle, i, at iteration k, relies on three components: (a) the velocity term in iteration k, $\mathbf{v}_i^k$, concerned by the constant inertia weight, ω; (b) the term called *cognitive learning*, which is the difference between the particle's best position found up until now (called $\mathbf{l}_i^k$, local best) and the particle's current position $\mathbf{x}_i^k$; and (c) the term of *social learning*, which is the difference between the best overall position found up to now in the whole swarm (called $\mathbf{g}^k$, global best) and the particle's current position $\mathbf{x}_i^k$. These two last terms are concerned in Eq. (13) by factors $\phi_1 = c_1 r_1$ and $\phi_2 = c_2 r_2$. In these two multipliers, c_1 and c_2 are constants, while r_1 and r_2 are random numbers distributed uniformly in the interval [0, 1]. Besides, in this study, the Standard PSO 2011 [45] has been utilized. It implies some improvements with respect to the preliminary implementations [21–24, 45]. Therefore, here, the PSO parameters are chosen as:

$$\omega = \frac{1}{2\ln 2} \quad \text{and} \quad c_1 = c_2 = 0.5 + \ln 2 \tag{15}$$

The swarm topology defines how the Np particles of the swarm are connected with each other to interchange information with the global best. In the actual Standard PSO, each particle informs only K particles, usually three chosen at random. A pure pseudo-code of the PSO algorithm is illustrated in Algorithm 1 below.

Algorithm 1. Pseudo-code of the PSO algorithm

Input: PSO population of particles $\mathrm{x}_i = \left(x_{i,1},\ldots,x_{i,n}\right)^T$ for $i = 1,\ldots,Np$; Np is the number of particles in the population

Output: The best solution g and its corresponding objective function value $f_{\min} = \min\left(f\left(\mathrm{x}\right)\right)$

```
initialize_particles;
$eval = 0$;
while termination_condition_not_met do
    for $i = 1$ to $Np$ do
        $f_i$ = evaluate_the_new_solution_objective_function$(\mathrm{x}_i)$;
        $eval = eval + 1$;
        if $f_i \le IBest_i$ then
            $\mathrm{I}_i = \mathrm{x}_i$; $IBest_i = f_i$; // save the local best solution
        end if
        if $f_i \le f_{\min}$ then
            $\mathrm{g} = \mathrm{x}_i$; $f_{\min} = f_i$; // save the global best solution
        end if
        $\mathrm{x}_i$ = generate_new_solution_with_equations_13_and_14$(\mathrm{x}_i)$;
    end for
end while
```

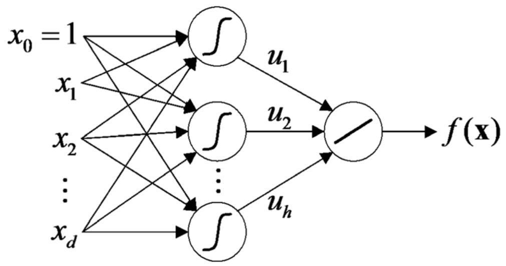

Fig. 5 Diagram of an MLP network with h neurons in the hidden layer, d neurons in the input layer, and a single neuron in the output layer

2.5 Artificial Neural Network: Multilayer Perceptron

Artificial neural networks (ANNs) are a computational model based on a large set of simple neuronal units (artificial neurons), roughly similar to the behavior observed in axons of neurons in biological brains [25–31]. The MLP is a kind of ANN made up of multiple layers that allows problems that are not linearly separable to be solved. Indeed, the MLP consists of an input layer and an output layer and one or more hidden layers of nonlinearly activating nodes [25, 26, 46]. It is a modification of the standard linear perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions (Fig. 5).

The MLP neural network introduces the function $\mathbf{f}: \mathbf{X} \subset \mathrm{R}^d \rightarrow \mathbf{Y} \subset \mathrm{R}^c$, which can be written as follows [25–31]:

$$\begin{aligned} \mathbf{f}(\mathbf{x}) &= \boldsymbol{\phi}(\boldsymbol{\psi}(\mathbf{x})) = (\boldsymbol{\phi} \circ \boldsymbol{\psi})(\mathbf{x}) \\ \boldsymbol{\phi} &: \mathbf{X} \subset \mathrm{R}^d \rightarrow \mathbf{U} \subset \mathrm{R}^h \\ \boldsymbol{\psi} &: \mathrm{U} \subset \mathrm{R}^h \rightarrow \mathbf{Y} \subset \mathrm{R}^c \end{aligned} \tag{16}$$

In Eq. (16), **U** is the space of hidden variables, termed the *characteristics space*. Relying on the established architecture, we have [25–31]:

- $\psi_j(\mathbf{x}) = \psi\left(\mathbf{w}_j^T \mathbf{x} + \mathbf{w}_{j0}\right)$: ψ is the activation function of the neurons of the hidden layer, $\mathbf{w}_j \in \mathfrak{R}^d$ is the vector of parameters of the different neurons, and $w_{j0} \in \mathfrak{R}$ is the threshold value. The three types of activation function ψ are sigmoid, logistic, and hyperbolic tangent.

Table 3 Set of physical-chemical input variables used in this study and their names along with their mean and standard deviation

Input variables	Name of the variable	Mean	Standard deviation
SO_2 (μg/m^3)	Sulfur dioxide	26.50	9.42
NO (μg/m^3)	Nitric oxide	68.32	12.47
NO_2 (μg/m^3)	Nitrogen dioxide	264.33	28.79
CO (mg/m^3)	Carbon monoxide	48.21	34.83
PM_{10} (μg/m^3)	Aerosol particles less than 10 μm	0.47	0.13
O_3 (μg/m3)	Ozone	1.24	0.23

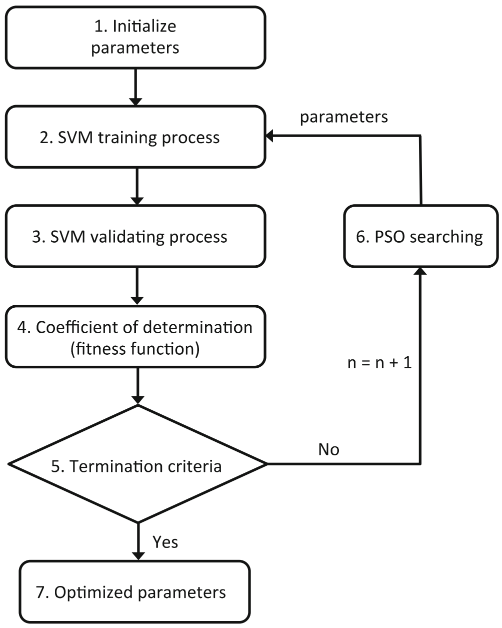

Fig. 6 Flowchart of the new hybrid PSO-RBF-SVM-based model

- $\phi_j(\mathbf{u}) = \phi\left(\mathbf{c}_j^T \mathbf{u} + \mathbf{c}_{j0}\right)$: ϕ is the activation function of the neurons of the output layer, $\mathbf{c}_j \in \mathfrak{R}^h$ is the vector of weights of the neurons, and $c_{j0} \in \mathfrak{R}$ is the threshold value. ϕ is normally the identity function, Heaviside function, or a dichotomous function.

The function concerned by the MLP is written as [25–31]:

$$\mathrm{f}(\mathbf{x}) = \sum_{j=1}^{h} c_j \psi\left(\mathbf{w}_j^T \mathbf{x} + w_{j0}\right) + c_0 \tag{17}$$

2.6 M5 Model Tree

The original algorithm M5 model tree was invented by Quinlan [32]. The M5 model tree (M5Tree) combines a conventional

Table 4 Optimal hyperparameters of the fitted PSO-RBF-SVM-based model found with the particle swarm optimization (PSO) technique for nitrogen dioxide (NO_2) in the metropolitan area of Oviedo

Kernel	Values of optimal hyperparameters
Linear	Regularization factor $C = 1.4206 \times 10^{-1}$, $\varepsilon = 6.0791 \times 10^{-2}$
Quadratic	Regularization factor $C = 6.8125 \times 10^{1}$, $\varepsilon = 2.9739 \times 10^{-6}$, $\sigma = 1.4150 \times 10^{-2}$, $a = 4.7847 \times 10^{-2}$, $b = 2$
Cubic	Regularization factor $C = 3.4989 \times 10^{1}$, $\varepsilon = 3.1132 \times 10^{-7}$, $\sigma = 5.1929 \times 10^{0}$, $a = 8.6717 \times 10^{-2}$, $b = 3$
Sigmoid	Regularization factor $C = 1.5314 \times 10^{-1}$, $\varepsilon = 1.3675 \times 10^{-5}$, $\sigma = 2.1554 \times 10^{-1}$, $a = 1.3682 \times 10^{-5}$
RBF	Regularization factor $C = 8.7135 \times 10^{0}$, $\varepsilon = 1.3221 \times 10^{-6}$, $\sigma = 3.2069 \times 10^{1}$

Table 5 The ANN parameters of the fitted multilayer perceptron (MLP) for nitrogen dioxide (NO_2) in the metropolitan area of Oviedo

Parameters	Values
Number of hidden neurons	9
Learning rate	0.1
Momentum factor	1.0×10^{-10}
Activation function	Tangent sigmoid transfer function

decision tree with the possibility of linear regression functions at the nodes (leaf) [33, 34]. The creation of the M5 model tree requires two different phases [46, 47]. During the first period, the dataset is divided into subsets so that a decision tree is built. The splitting criterion uses the standard deviation of the class values and the expected lowering in this error. The standard deviation reduction (SDR) can be calculated as [32–34, 48]:

$$\mathrm{SDR} = \mathrm{sd}(T) - \sum \frac{|T_i|}{|T|}\mathrm{sd}(T_i) \quad (18)$$

where T is the set of instances that reach this node, T_i are the sets that result from splitting the node according to the chosen attribute, and sd is the standard deviation of the class values [47, 48]. The splitting process finishes when the class values of the instances that reach a node vary only slightly, that is to say, when their standard deviation is only a small fraction (for instance, less than 5%) of the standard deviation of the original instance set. As a result of the splitting process, the data on the secondary nodes have less standard deviation compared to the parent nodes and thus are purer children. M5Tree chooses the one that maximizes the expected error reduction after scanning all possible divisions. This splitting often gives rise to an extremely large tree-like structure and may produce unsatisfactory performance. To address this problem, the huge tree is pruned and the nodes of the tree are substituted by linear regression functions in the second phase [49].

Table 6 Coefficient of determination (R^2) and correlation coefficient (r) for the hybrid PSO-SVM-based models (with linear, quadratic, cubic, sigmoid, and RBF kernels), multilayer perceptron (MLP) approach, and M5 tree model fitted in this study for nitrogen dioxide (NO_2) in the metropolitan area of Oviedo

Model	Coefficients of determination (R^2)/correlation coefficients (r)
Linear-SVM	0.71/0.84
Quadratic-SVM	0.74/0.86
Cubic-SVM	0.79/0.89
Sigmoid-SVM	0.71/0.84
RBF-SVM	0.98/0.99
Multilayer perceptron	0.76/0.87
M5 model tree	0.79/0.89

Table 7 Weights of the variables in the fitted PSO-RBF-SVM-based model for the nitrogen dioxide (NO_2) value in the metropolitan area of Oviedo

Variable	Weight
Nitric oxide (NO)	4.6232
Sulfur dioxide (SO_2)	2.7424
Ozone (O_3)	− 2.7399
Carbon monoxide (CO)	0.0172
Aerosol particles less than 10 μm (PM_{10})	− 0.0426

3 Results and Discussion

The physical-chemical input variables taken into account in this research are shown in Table 3 [1–12, 35–37, 40–42]. The total number of predicting variables used to carry out the regression of the hybrid PSO-SVM-based model, MLP approach, and M5 model tree was 5. Besides, the total number of output-predicted dependent variables was 3: NO_2, SO_2, and PM_{10}. Indeed, we have constructed three different models taking as dependent variables NO_2, SO_2, and PM_{10}, respectively. Additionally, as independent input variables (predictor variables), the other remaining variables listed in Table 3 were also considered.

On the one hand, the SVM techniques are very dependent on the values of their hyperparameters. Also, the number of hyperparameters relies on the type of kernel chosen. Among these, we can mention: the regularization factor C (Eq. 4), the value of ε that defines the width of the insensitive tube (permitted error), and the remaining hyperparameters commonly called a, b, and σ. For instance, grid search, genetic algorithms, and artificial bee colony (ABC) are optimization methods habitually used to determine the appropriate SVR parameters of each kernel [19, 20]. The grid search method used by most computational codes is a brute force method, and as such, almost any optimization

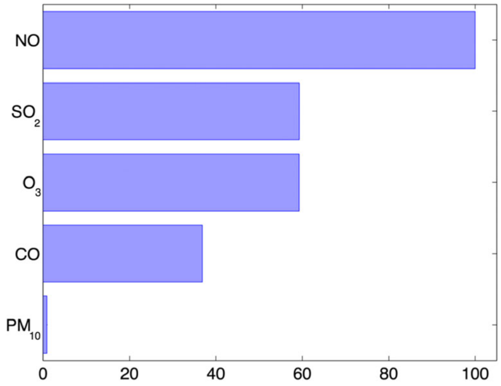

Fig. 7 Relative importance of the input variables to predict the nitrogen dioxide (NO_2) value in the metropolitan area of Oviedo in the fitted PSO-RBF-SVM-based model

Table 8 Optimal hyperparameters of the fitted PSO-RBF-SVM-based model found with the particle swarm optimization (PSO) technique for sulfur dioxide (SO_2) in the metropolitan area of Oviedo

Kernel	Values of optimal hyperparameters
Linear	Regularization factor $C = 1.8911 \times 10^0$, $\varepsilon = 8.9098 \times 10^{-2}$
Quadratic	Regularization factor $C = 3.5705 \times 10^1$, $\varepsilon = 8.2453 \times 10^{-2}$, $\sigma = 2.7744 \times 10^{-1}$, $a = 2.7525 \times 10^{-2}$, $b = 2$
Cubic	Regularization factor $C = 2.3067 \times 10^1$, $\varepsilon = 2.6629 \times 10^{-2}$, $\sigma = 1.6104 \times 10^0$, $a = 7.0757 \times 10^{-1}$, $b = 3$
Sigmoid	Regularization factor $C = 8.6787 \times 10^1$, $\varepsilon = 9.0458 \times 10^{-2}$, $\sigma = 2.0674 \times 10^{-3}$, $a = 2.1430 \times 10^{-3}$
RBF	Regularization factor $C = 3.1955 \times 10^0$, $\varepsilon = 3.1654 \times 10^{-8}$, $\sigma = 3.9841 \times 10^1$

Table 10 Coefficient of determination (R^2) and correlation coefficient (r) for the hybrid PSO-SVM-based models (with linear, superlinear, quadratic, cubic, sigmoid, and RBF kernels), multilayer perceptron (MLP) approach, and M5 tree model fitted in this study for sulfur dioxide (SO_2) in the metropolitan area of Oviedo

Model	Coefficients of determination (R^2)/correlation coefficients (r)
Linear-SVM	0.50/0.71
Quadratic-SVM	0.64/0.80
Cubic-SVM	0.69/0.83
Sigmoid-SVM	0.50/0.71
RBF-SVM	0.94/0.97
Multilayer perceptron	0.67/0.82
M5 model tree	0.65/0.81

method improves its efficiency. Specifically, in this study, we have utilized the PSO optimization technique [21–24] for tuning the SVR parameters so that a hybrid PSO-SVM-based model was fitted to experimental dataset to predict the output-dependent variables (NO_2, SO_2, and PM_{10}) from the other remaining variables (input variables) in an air quality analysis [35–37, 40–42] with success. As a statistical estimator of the goodness of fit, the coefficient of determination R^2 was used successfully. Figure 6 depicts the flowchart of this new hybrid PSO-SVM-based model implemented in this study.

If we now apply the PSO technique, the so-called particles $\mathbf{x}_i$ include the tuning parameters. For instance, if we choose the RBF as the kernel, then the components of the particle are written as $\mathbf{x}_i = (C_i, \varepsilon_i, \sigma_i)$. According to the PSO algorithm, we randomly initialize these parameters in the first stage. For the next iterations, the particles evolve following Eqs. (13) and (14). Then, the objective function value for all the particles is determined in each iteration. Specifically, the objective function value is the minus tenfold cross-validation coefficient of determination for each particle. If the termination criteria are satisfied, the global best $\mathbf{x}_i$ contains the optimized parameters. Therefore, tenfold cross-validation was the standard technique used here for finding the real coefficient of determination (R^2) [50–53]. The combination of the hyperparameters with the best efficiency is termed *optimal hyperparameters* [13–20, 52, 53].

The support vector regression has been carried out with the SVR-ε method using the LIBSVM library [54], and the hyperparameters have been optimized with PSO, utilizing the standard PSO 2011 version [45, 55, 56]. The searching in the parameter space has been done taking into account that the SVM algorithm significantly changes its results when its parameters increase or decrease in a power of 10. For instance, in the case of RBF kernel, we have considered [−6, 2] × [−10, 2] × [−6, 2]. That is, C values (regularization parameter) varies within the interval $[10^{-6}, 10^2]$, ε values within $[10^{-10}, 10^2]$, and σ values within $[10^{-6}, 10^2]$ in the optimization stage. The stopping criterion is met if there is no improvement in the R^2 after ten iterations, in combination with a maximum number of iterations equal to 500.

Table 4 shows the optimal hyperparameters of the fitted PSO-RBF-SVM-based model found with the PSO technique for NO_2 in the metropolitan area of Oviedo on a local scale.

An iMac with a 3.2-GHz Intel Core i5 CPU with 8 Gb of RAM and Mavericks as operating system was used. The stopping conditions, ten iterations without improvement or a maximum of 300 iterations, were met after 75 iterations and 4 h and 22 min for NO_2.

Similarly, and for purposes of comparison, a MLP and M5 tree model have been fitted to the experimental data corresponding to NO_2 in order to predict its value in the metropolitan area of Oviedo on a local scale. In this sense, an ANN is typically defined by three types of parameters [25–31]: the

Table 9 The ANN parameters of the fitted multilayer perceptron (MLP) for sulfur dioxide (SO_2) in the metropolitan area of Oviedo

Parameters	Values
Number of hidden neurons	9
Learning rate	0.1
Momentum factor	0.001
Activation function	Tangent sigmoid transfer function

Table 11 Weights of the variables in the fitted PSO-RBF-SVM-based model for the sulfur dioxide (SO_2) value in the metropolitan area of Oviedo

Variable	Weight
Nitric oxide (NO)	3.3559
Nitrogen dioxide (NO_2)	2.5972
Carbon monoxide (CO)	1.6660
Ozone (O_3)	− 1.6075
Aerosol particles less than 10 μm (PM_{10})	1.4966

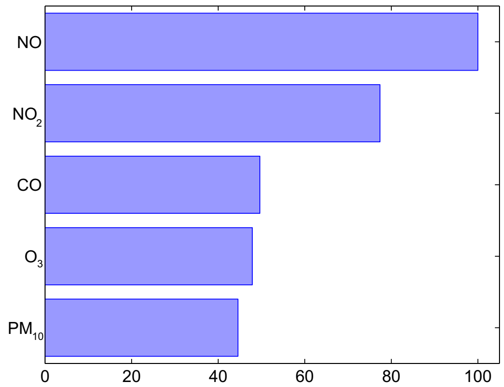

Fig. 8 Relative importance of the input variables to predict the sulfur dioxide (SO_2) value in the metropolitan area of Oviedo in the fitted PSO-RBF-SVM-based model

interconnection pattern between different layers of neurons (Fig. 5), the learning process for updating the weights of the interconnections, the momentum factor in order to avoid oscillating weight changes of the ANN, and the activation function that converts a neuron's weighted input to its output activation. In this paper, the ANN optimal parameters for the MLP are shown in Table 5.

Table 6 shows the determination and correlation coefficients for the PSO-SVM-based models for the five kernels (linear, quadratic, cubic, sigmoid, and RBF kernels, respectively), multilayer perceptron, and M5 tree model fitted here for NO_2 in the metropolitan area of Oviedo.

According to the statistical calculations, the SVM with the RBF kernel function is the best model for estimating the concentration of NO_2 in the metropolitan area of Oviedo on a local scale, since the fitted SVM with RBF kernel function has a coefficient of determination R^2 equal to 0.9802 and a correlation coefficient equal to 0.9900. These results indicate an important goodness of fit, that is to say, a very good agreement is obtained between our model and the observed data. Furthermore, the importance ranking of the five remaining input variables (Table 3) in order to predict the NO_2 value (output variable) in this nonlinear complex problem is shown in Table 7 and Fig. 7.

Following the same methodology, fittings were also made for SO_2 and PM_{10} as dependent variables using the PSO-RBF-SVM-based model, MLP technique, and M5 model tree, whose results we show below. Indeed, Table 8 shows the optimal hyperparameters of the fitted PSO-RBF-SVM-based model found with the PSO technique for SO_2 in the metropolitan area of Oviedo on a local scale.

The stopping conditions, ten iterations without improvement or a maximum of 300 iterations, were met after 84 iterations and 5 h and 41 min for SO_2.

In this paper, the ANN parameters of the fitted MLP for SO_2 in the metropolitan area of Oviedo are shown in Table 9.

Similarly, Table 10 shows the determination and correlation coefficients for the PSO-SVM-based models for the five kernels (linear, quadratic, cubic, sigmoid, and RBF kernels, respectively), multilayer perceptron, and M5 tree model fitted here for SO_2 in the metropolitan area of Oviedo.

According to the statistical calculations, the SVM with the RBF kernel function is the best model for estimating the concentration of NO_2 in the metropolitan area of Oviedo on a local scale, since the fitted SVM with RBF kernel function has a coefficient of determination R^2 equal to 0.9499 and a correlation coefficient equal to 0.9746. These results indicate an important goodness of fit, that is to say, a very good agreement is obtained between our model and the observed data. Furthermore, the importance ranking of the five remaining input variables (Table 3) in order to predict the NO_2 value (output variable) in this nonlinear complex problem is shown in Table 11 and Fig. 8.

Next, Table 12 shows the optimal hyperparameters of the fitted PSO-RBF-SVM-based model found with the PSO technique for PM_{10} in the metropolitan area of Oviedo on a local scale.

The stopping conditions, ten iterations without improvement or a maximum of 300 iterations, were met after 69 iterations and 4 h and 6 min for PM_{10}.

In this paper, the ANN parameters of the fitted MLP for aerosol less than 10 μm (PM_{10}) in the metropolitan area of Oviedo are shown in Table 13.

Similarly, Table 14 shows the determination and correlation coefficients for the PSO-SVM-based models for the five kernels (linear, quadratic, cubic, sigmoid, and RBF kernels, respectively), multilayer perceptron, and M5 tree model fitted here for aerosol less than 10 μm (PM_{10}) in the metropolitan area of Oviedo.

Table 12 Optimal hyperparameters of the fitted PSO-RBF-SVM-based model found with the particle swarm optimization (PSO) technique for particulate matter less than 10 μm (PM_{10}) in the metropolitan area of Oviedo

Kernel	Values of optimal hyperparameters
Linear	Regularization factor $C = 1.3130 \times 10^{-1}$, $\varepsilon = 1.3493 \times 10^{-1}$
Quadratic	Regularization factor $C = 6.6965 \times 10^{1}$, $\varepsilon = 8.0252 \times 10^{-2}$, $\sigma = 1.2451 \times 10^{0}$, $a = 7.5358 \times 10^{-1}$, $b = 2$
Cubic	Regularization factor $C = 8.8805 \times 10^{2}$, $\varepsilon = 5.7101 \times 10^{-2}$, $\sigma = 3.2484 \times 10^{-1}$, $a = 9.6975 \times 10^{-1}$, $b = 3$
Sigmoid	Regularization factor $C = 3.7271 \times 10^{1}$, $\varepsilon = 1.2589 \times 10^{-7}$, $\sigma = 2.6951 \times 10^{-3}$, $a = 1.1634 \times 10^{-5}$
RBF	Regularization factor $C = 3.7120 \times 10^{0}$, $\varepsilon = 1.1093 \times 10^{-9}$, $\sigma = 5.7759 \times 10^{1}$

Table 13 The ANN parameters of the fitted multilayer perceptron (MLP) for aerosol less than 10 μm (PM_{10}) in the metropolitan area of Oviedo

Parameters	Values
Number of hidden neurons	11
Learning rate	0.1
Momentum factor	0.1
Activation function	Tangent sigmoid transfer function

Additionally, according to the statistical calculations, the SVM with the RBF kernel function is the best model for estimating the concentration of PM_{10} in the metropolitan area of Oviedo on a local scale, since the fitted SVM with RBF kernel function has a coefficient of determination R^2 equal to 0.8458 and a correlation coefficient equal to 0.9197.

Furthermore, the importance ranking of the five remaining input variables (Table 3) in order to predict PM_{10} value (output variable) in this nonlinear complex problem is shown in Table 15 and Fig. 9.

From the results depicted in Table 7 and Fig. 7, it is possible to observe that the most important variables for the prediction of the NO_2 (output variable) according to the PSO-RBF-SVM model are in hierarchical order: NO, SO_2, O_3, CO, and PM_{10}. The influence of the variable PM_{10} was negligible, according to the calculations. The most significant variable in NO_2 prediction is NO. This result is logical since NO_2 typically arises via the oxidation of NO by oxygen in air. Nitrogen dioxide is formed in most combustion processes using air as the oxidant.

Similarly, the results shown in Table 11 and Fig. 8 indicate that the most important variables for the prediction of SO_2 (output variable) are NO, NO_2, CO, O_3, and PM_{10}. Again, the influence of the variable PM_{10} was the smallest, according to the calculations. SO_2 is the product of the burning of sulfur or of burning materials that contain sulfur. Furthermore, sulfur dioxide emissions are a precursor to acid rain and atmospheric particulates.

Table 14 Coefficient of determination (R^2) and correlation coefficient (r) for the hybrid PSO-SVM-based models (with linear, quadratic, cubic, sigmoid, and RBF kernels), multilayer perceptron (MLP) approach, and M5 tree model fitted in this study for aerosol less than 10 μm (PM_{10}) in the metropolitan area of Oviedo

Model	Coefficients of determination (R^2)/correlation coefficients (r)
Linear-SVM	0.12/0.35
Quadratic-SVM	0.20/0.45
Cubic-SVM	0.36/0.60
Sigmoid-SVM	0.11/0.33
RBF-SVM	0.85/0.92
Multilayer perceptron	0.32/0.57
M5 model tree	0.33/0.58

Table 15 Weights of the variables in the fitted PSO-RBF-SVM-based model for the particulate matter less than 10 μm (PM_{10}) value in the metropolitan area of Oviedo

Variable	Weight
Nitric oxide (NO)	2.2389
Sulfur dioxide (SO_2)	1.7814
Carbon monoxide (CO)	1.4004
Nitrogen dioxide (NO_2)	0.6798
Ozone (O_3)	− 0.5785

From the results shown in Table 15 and Fig. 9, the most important variables for the prediction of PM_{10} (output variable) are NO, SO_2, CO, NO_2, and O_3. The influence of the variables NO_2 and O_3 was negligible, according to the calculations. Some particulates occur naturally, originating from volcanoes, dust storms, forest and grassland fires, living vegetation, and sea spray. Human activities, such as the burning of fossil fuels in vehicles, power plants, and various industrial processes, also generate significant amounts of particulates (anthropogenic aerosols). In this way, secondary particles are derived from the oxidation of primary gases such as sulfur and nitrogen oxides into sulfuric acid (liquid) and nitric acid (gaseous). The precursors for these aerosols (i.e., the gases from which they originate) may have an anthropogenic origin (from fossil fuel or coal combustion) and a natural biogenic origin.

Finally, this research allows the prediction of the concentrations of NO_2 from 2013 to 2015 in agreement with the actual experimental concentrations of NO_2 observed using the PSO-RBF-SVM-based model with great accuracy and success. Indeed, Fig. 10 shows the comparison between the NO_2 values observed and predicted by using the M5 model tree (Fig. 10a), MLP (Fig. 10b), and PSO-SVM-based model with RBF kernel (Fig. 10c). It is necessary to use a SVM model with RBF kernel in order to achieve the best effective approach to nonlinearities present in this regression problem. Obviously, these results again coincide with the outcome

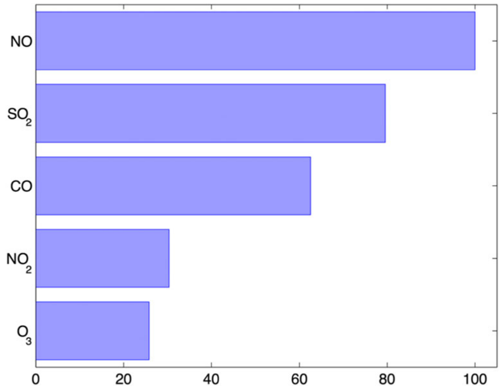

Fig. 9 Relative importance of the input variables to predict the sulfur dioxide (SO_2) value in the metropolitan area of Oviedo in the fitted PSO-RBF-SVM-based model

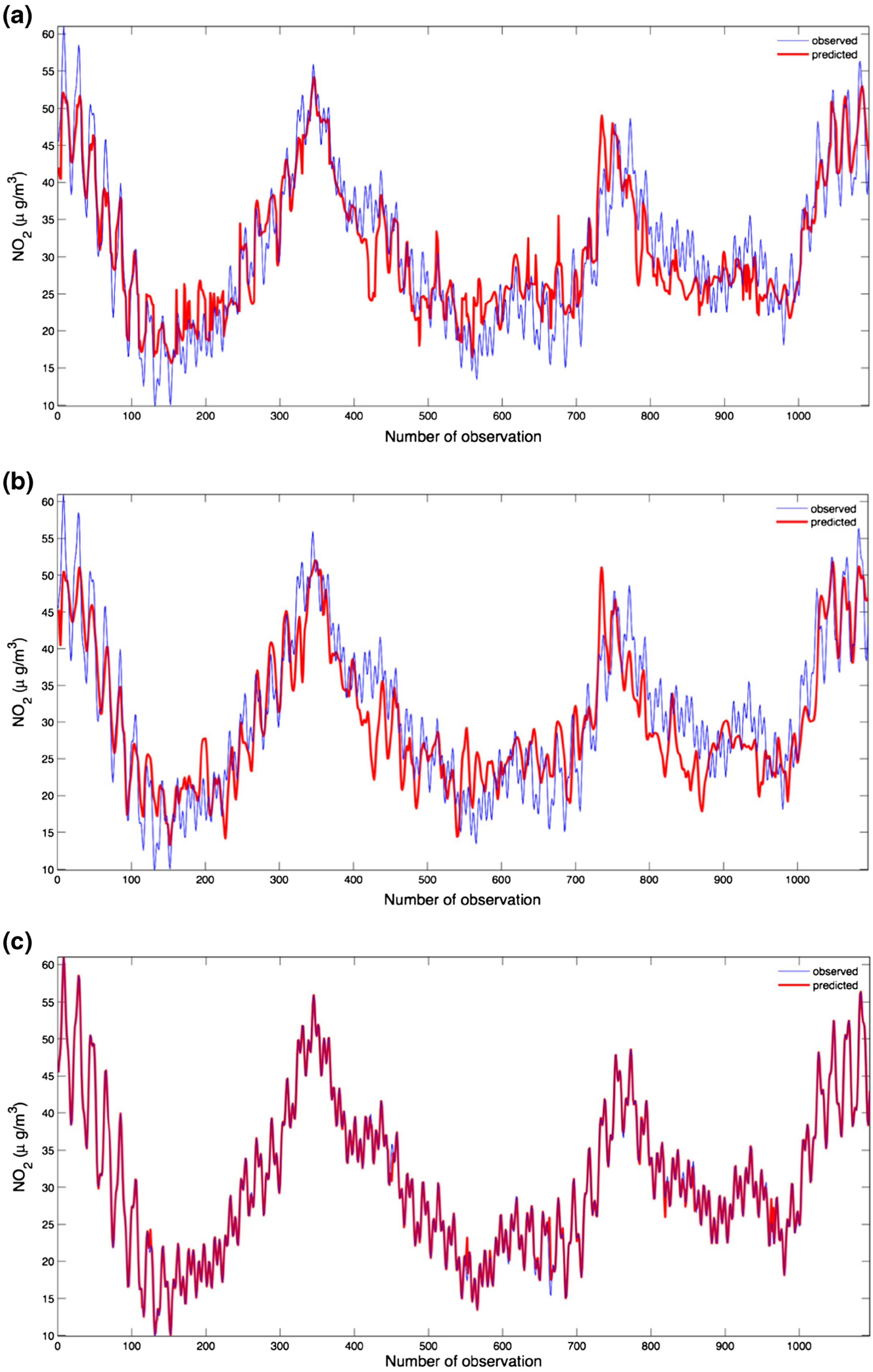

Fig. 10 Comparison between NO_2 values observed and predicted by the M5 model tree, the MLP approach, and the PSO-SVM-based model: **a** M5 model tree ($R^2 = 0.75$), **b** MLP network ($R^2 = 0.80$), and **c** PSO-SVM model with RBF kernel ($R^2 = 0.9802$)

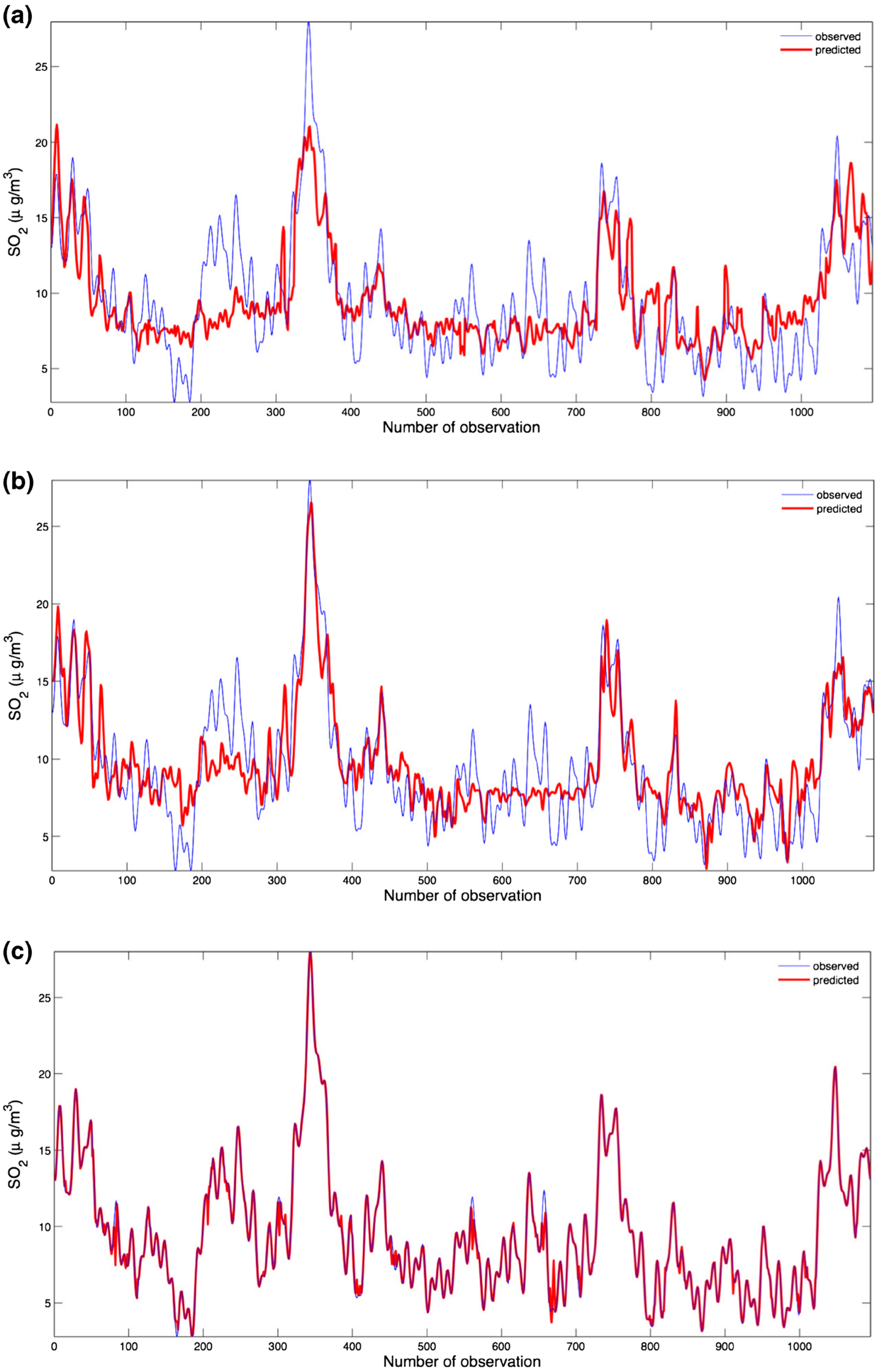

Fig. 11 Comparison between SO_2 values observed and predicted by the M5 model tree, the MLP approach, and the PSO-SVM-based model: **a** M5 model tree ($R^2 = 0.75$), **b** MLP network ($R^2 = 0.80$), and **c** PSO-SVM model with RBF kernel ($R^2 = 0.9499$)

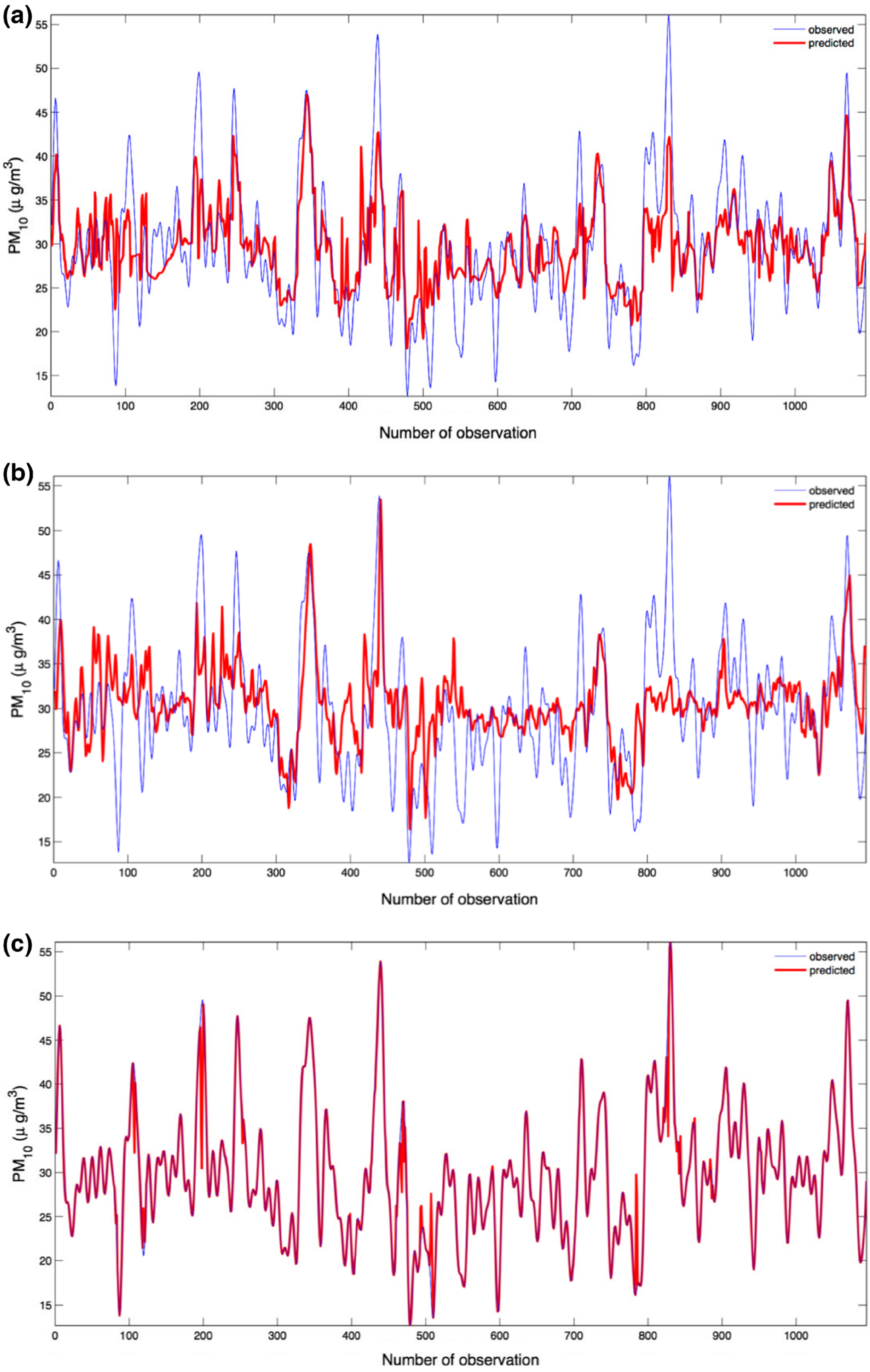

Fig. 12 Comparison between PM_{10} concentrations observed and predicted by the M5 model tree, the MLP approach, and the PSO-SVM-based model: **a** M5 model tree ($R^2 = 0.75$), **b** MLP network ($R^2 = 0.80$), and **c** PSO-SVM model with RBF kernel ($R^2 = 0.8458$)

criterion of 'goodness of fit' (R^2) so that the PSO-SVM-based model with a RBF kernel function was the best fitting.

Similarly, this study was also able to predict the concentrations of SO_2 and PM_{10} from 2013 to 2015 in agreement with the actual experimental concentrations of SO_2 and PM_{10} observed using the PSO-RBF-SVM-based model with great accuracy and success. Indeed, Figs. 11 and 12 below show the comparison between the SO_2 and PM_{10} values observed and predicted by using the M5 model tree (Figs. 11a and 12a), MLP (Figs. 11b and 12b), and PSO-SVM-based model with RBF kernel (Figs. 11c and 12c), respectively. It is mandatory to use a SVM model with RBF kernel in order to achieve the best effective approach to nonlinearities present in this regression problem. Obviously, these results again coincide with the outcome criterion of 'goodness of fit' (R^2) so that the PSO-SVM-based model with a RBF kernel function was the best fitting.

4 Conclusions

Despite widespread success, the challenges to air quality management remain completely unresolved today. Based on the experimental and numerical results, the main findings of this research work can be summarized as follows:

- Firstly, all governments have announced plans for improving air quality in cities while minimizing the impact on business. However, emission reduction strategies to avoid litigation and satisfy the public and other stakeholders are very difficult to carry out in practice, requiring perhaps years of implementation. Furthermore, the diagnostic techniques commonly used based on the traditional methods (e.g., monitoring of pollutants through automatic stations) are expensive, from both the material and human points of view. Consequently, the development of alternative diagnostic techniques is necessary. In this sense, the new hybrid PSO-SVM-based method with a RBF kernel function used in this research is a very good choice for evaluating the air quality in cities on a local scale.
- Secondly, the hypothesis was confirmed that air quality diagnosis in the metropolitan area of Oviedo can be accurately modeled by using a hybrid PSO-SVM-based model with a RBF kernel function on a local scale.
- Thirdly, a hybrid PSO-SVM-based model with a RBF kernel function was successfully developed to predict the concentrations of NO_2, SO_2, and PM_{10} from the other measured input operation pollutants, in order to lower costs in the assessment of air quality in the metropolitan area of Oviedo.
- Fourthly, high coefficients of determination equal to 0.9802, 0.9499, and 0.8458 were obtained when this hybrid PSO-SVM-based model with a RBF kernel function was applied to the experimental dataset corresponding to pollutants in the metropolitan area of Oviedo. Indeed, the predicted results for this model have proven to be consistent with the historical dataset of actual observed values of the pollutants from 2013 to 2015 (Figs. 10, 11, and 12).
- Fifthly, the order of significance of the input variables involved in the prediction of the concentrations of NO_2, SO_2, and PM_{10} was set. This is one of the main findings in this study.
- Sixthly, the influence of the kernel parameters setting of the SVMs on the regression performance of the value of the air quality was established.
- Finally, the results of this research concerning the development of models of local pollutant concentrations will prove to be a valuable tool for projects on the mitigation of acid rain and for the research into the effects of particulate matter on human health. Furthermore, there is an increasing interest in the use of mathematical models with good physical properties to better understand the behavior of the pollutants in the atmosphere so as to improve the air quality and reduce the number of deaths. The results verify that the hybrid PSO-SVM regression method significantly improves the generalization capability achievable with only the SVM-based regressor. In this sense, this model can be assembled inside other, more general models of the atmosphere.

In summary, this innovative methodology could be successfully applied to other cities or locations with similar or different types of pollutants, but it is always mandatory to take into account the specificities of each place. Consequently, an effective PSO-SVM-based model is a practical solution to the problem of the determination of the air quality in cities. This methodology allows areas of each city where the air quality problem is less serious to be labeled as clean air zones. Furthermore, this paper presents examples of real applications and simple explanations of statistical calculation for the selection of the best-fitted models.

Acknowledgements The authors wish to acknowledge the computational support provided by the Department of Mathematics at University of Oviedo. Additionally, we would like to thank Anthony Ashworth for his revision of the English grammar and spelling of the manuscript.

References

1. García Nieto, P. J. (2001). Parametric study of selective removal of atmospheric aerosol by coagulation, condensation and gravitational settling. *International Journal of Environmental Health Research, 11*, 151–162.
2. García Nieto, P. J. (2006). Study of the evolution of aerosol emissions from coal-fired power plants due to coagulation, condensation, and gravitational settling and health impact. *Journal of Environmental Management, 79*(4), 372–382.

3. Lutgens, F. K., & Tarbuck, E. J. (2001). *The atmosphere: an introduction to meteorology.* New York: Prentice Hall.
4. Wark, K., Warner, C. F., & Davis, W. T. (1997). *Air pollution: its origin and control.* New Jersey: Prentice Hall.
5. Wang, L. K., Pereira, N. C., & Hung, Y. T. (2004). *Air pollution control engineering*. New York: Humana.
6. Karaca, F., Alagha, O., & Ertürk, F. (2005). Statistical characterization of atmospheric PM_{10} and $PM_{2.5}$ concentrations at a non-impacted suburban site of Istanbul, Turkey. *Chemosphere, 59*(8), 1183–1190.
7. Comrie, A. C., & Diem, J. E. (1999). Climatology and forecast modeling of ambient carbon monoxide in Phoenix. *Atmospheric Environment, 33*, 5023–5036.
8. Elbir, T., & Muezzinoglu, A. (2000). Evaluation of some air pollution indicators in Turkey. *Environment International, 26*(1–2), 5–10.
9. Godish, T., Davis, W. T., & Fu, J. S. (2014). *Air quality.* Boca Raton: CRC.
10. Akkoyunku, A., & Ertürk, F. A. (2003). Evaluation of air pollution trends in Istanbul. *International Journal of Environment and Pollution, 18*, 388–398.
11. Suárez Sánchez, A., García Nieto, P. J., Riesgo Fernández, P., del Coz Díaz, J. J., & Iglesias-Rodríguez, F. J. (2011). Application of a SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). *Mathematical and Computer Modelling, 54*(5–6), 1453–1466.
12. Cooper, C. D., & Alley, F. C. (2002). *Air pollution control*. New York: Waveland Press.
13. Vapnik, V. (1998). *Statistical learning theory.* New York: Wiley-Interscience.
14. Hastie, T., Tibshirani, R., & Friedman, J. (2003). *The elements of statistical learning*. New York: Springer.
15. Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.
16. Schölkopf, B., Smola, A. J., Williamson, R., & Bartlett, P. (2000). New support vector algorithms. *Neural Computing and Applications, 12*(5), 1207–1245.
17. Hansen, T., & Wang, C. J. (2005). Support vector based battery state of charge estimator. *Journal of Power Sources, 141*, 351–358.
18. Li, X., Lord, D., Zhang, Y., & Xie, Y. (2008). Predicting motor vehicle crashes using support vector machine models. *Accident Analysis & Prevention, 40*, 1611–1618.
19. Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. New York: Cambridge University Press.
20. Steinwart, I., & Christmann, A. (2008). *Support vector machines*. New York: Springer.
21. Kennedy, J., & Eberhart, R. C. (1995). *Particle swarm optimization. In Proceedings of the fourth IEEE international conference on neural networks* (Vol. 4, pp. 1942–1948). Perth: IEEE Service Center.
22. Eberhart, R. C., Shi, Y., & Kennedy, J. (2001). *Swarm intelligence*. San Francisco: Morgan Kaufmann.
23. Clerc, M. (2006). *Particle swarm optimization*. London: Wiley-ISTE.
24. Olsson, A. E. (2011). *Particle swarm optimization: theory, techniques and applications*. New York: Nova Science Publishers.
25. Boznar, M., Lesjack, M., & Mlakar, P. (1993). A neural network based method for short-term predictions of ambient SO_2 concentrations in highly polluted industrial areas of complex terrain. *Atmospheric Environment, 270*, 221–230.
26. Haykin, S. (1999). *Neural networks: comprehensive foundation*. New Jersey: Prentice Hall.
27. Hooyberghs, J., Mensink, C., Dumont, D., Fierens, F., & Brasseur, O. (2005). A neural network forecast for daily average PM_{10} concentrations in Belgium. *Atmospheric Environment, 39*(18), 3279–3289.
28. Kukkonen, J., Partanen, L., Karpinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., & Cawley, G. (2003). Extensive evaluation of neural networks models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmospheric Environment, 37*, 4539–4550.
29. Gardner, M. W., & Dorling, S. R. (1999). Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. *Atmospheric Environment, 33*(5), 709–719.
30. Chaloulakou, A., Saisana, M., & Spyrellis, N. (2003). Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. *Science of the Total Environment, 313*, 1–13.
31. Karaca, F., Nikov, A., & Alagha, O. (2006). NN-AirPol: a neural-network-based method for air pollution evaluation and control. *International Journal of Environment and Pollution, 28*(3–4), 310–325.
32. Quinlan, J. R. (1992). Learning with continuous classes. In *Proceedings of Australian joint conference on artificial intelligence* (pp. 343–348). Singapore: World Scientific Press.
33. Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Monterrey: Wadsworth and Brooks.
34. Kisi, O. (2015). Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *Journal of Hydrology, 528*, 312–320.
35. Colbeck, I. (2008). *Environmental chemistry of aerosol*. New York: Wiley-Blackwell.
36. Hewitt, C. N., & Jackson, A. V. (2009). *Atmospheric science for environmental scientists*. New York: Wiley-Blackwell.
37. Schnelle, K. B., Dunn, R. F., & Ternes, M. E. (2015). *Air pollution control technology handbook*. Boca Raton: CRC.
38. Simon, D. (2013). *Evolutionary optimization algorithms*. New York: Wiley.
39. Yang, X.-S., Cui, Z., Xiao, R., Gandomi, A. H., & Karamanoglu, M. (2013). *Swarm intelligence and bio-inspired computation: theory and applications*. London: Elsevier.
40. Monteiro, A., Lopes, M., Miranda, A. I., Borrego, C., & Vautard, R. (2005). Air pollution forecast in Portugal: a demand from the new air quality framework directive. *International Journal of Environment and Pollution, 5*, 1–9.
41. Friedlander, S. K. (2000). *Smoke, dust and haze: fundamentals of aerosol dynamics*. New York: Oxford University Press.
42. Vincent, J. H. (2007). *Aerosol sampling: science, standards, instrumentation and applications*. Chichester: Wiley.
43. de Cos Juez, F. J., García Nieto, P. J., Martínez Torres, J., & Taboada Castro, J. (2010). Analysis of lead times of metallic components in the aerospace industry through a supported vector machine model. *Mathematical and Computer Modelling, 52*, 1177–1184.
44. Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. New York: Cambridge University Press.
45. Clerc, M. (2012). *Standard particle swarm optimisation: from 2006 to 2011*. Technical report. http://clerc.maurice.free.fr/pso/SPSO_descriptions.pdf. Accessed 23 Sept 2012.
46. Solomatine, D. P., & Xue, Y. P. (2004). M5 model trees and neural networks: application to flood forecasting in the upper reach of the Hual River in China. *Journal of Hydrologic Engineering, 9*(6), 491–501.
47. Rahimikhoob, A., Asadi, M., & Mashal, M. (2013). A comparison between conventional and M5 model tree methods for converting pan evaporation to reference evapotranspiration for semi-arid region. *Water Resources Management, 27*(14), 4815–4826.
48. Pal, M., & Deswal, S. (2009). M5 model tree based modelling of reference evapotranspiration. *Hydrological Processes, 23*(10), 1437–1443.

49. Pal, M. (2006). M5 model tree for land cover classification. *International Journal of Remote Sensing, 27*(4), 825–831.
50. Wasserman, L. (2003). *All of statistics: a concise course in statistical inference*. New York: Springer.
51. Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics*. New York: W.W. Norton & Company.
52. Picard, R., & Cook, D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association, 79*(387), 575–583.
53. Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: the .632 + bootstrap method. *Journal of the American Statistical Association, 92*(438), 548–560.
54. Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*, 1–27.
55. Zambrano-Bigiarini, M., & Rojas, R. (2013). A model-independent particle swarm optimisation software for model calibration. *Environmental Modelling & Software, 43*, 5–25.
56. Zambrano-Bigiarini, M., & Rojas, R. (2014). *HydroPSO: a flexible and model-independent particle swarm optimisation (PSO) package for calibration/optimisation of environmental models. In R package, version 0.3–4*. Vienna: R Foundation for Statistical Computing.