Science of the Total Environment 621 (2018) 753–761

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

PM_{10} concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: A case study

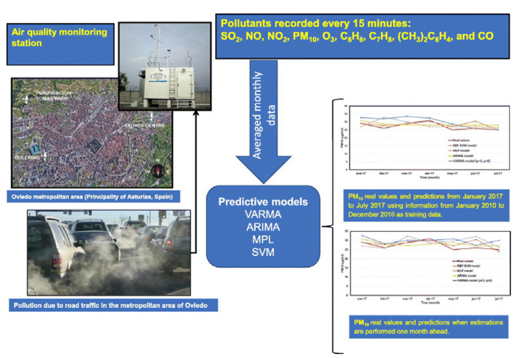
P.J. García Nieto [a,*], F. Sánchez Lasheras [a], E. García-Gonzalo [a], F.J. de Cos Juez [b]

[a] Department of Mathematics, Faculty of Sciences, University of Oviedo, 33007 Oviedo, Spain
[b] Exploitation and Prospecting Department, University of Oviedo, 33004 Oviedo, Spain

HIGHLIGHTS

- Four models based on SVM, MLP, VARMA and ARIMA are built for forecasting of the PM10 concentration in the city of Oviedo.
- PM_{10} have impacts on climate and precipitation that adversely affect human health.
- The description of the air quality is of real interest for the effective safety management of the air pollution in cities.
- The results show that the SVM model was better than the other models to forecast PM_{10} concentration.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 24 October 2017
Received in revised form 25 November 2017
Accepted 26 November 2017
Available online xxxx

Editor: Jianmin Chen

Keywords:
Particulate matter (PM_{10}) forecasting
Support vector regression (SVR)
Multilayer perceptron (MLP)
Vector autoregressive moving-average (VARMA)
Autoregressive integrated moving-average (ARIMA)

ABSTRACT

Atmospheric particulate matter (PM) is one of the pollutants that may have a significant impact on human health. Data collected over seven years in a city of the north of Spain is analyzed using four different mathematical models: vector autoregressive moving-average (VARMA), autoregressive integrated moving-average (ARIMA), multilayer perceptron (MLP) neural networks and support vector machines (SVMs) with regression. Measured monthly average pollutants and PM_{10} (particles with a diameter less than 10 μm) concentration are used as input to forecast the monthly averaged concentration of PM_{10} from one to seven months ahead. Simulations showed that the SVM model performs better than the other models when forecasting one month ahead and also for the following seven months.

1. Introduction

Atmospheric aerosol particles, also known as *atmospheric particulate matter* (PM), are microscopic solid or liquid matter suspended in Earth's atmosphere. Sources of particulate matter can be *natural* or

* Corresponding author.
E-mail address: lato@orion.ciencias.uniovi.es (P.J. García Nieto).

https://doi.org/10.1016/j.scitotenv.2017.11.291
0048-9697/

anthropogenic (Friedlander, 2000). 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. Coal combustion in developing countries is the primary method for heating homes and supplying energy. Anthropogenic aerosols (those made by human activities) currently account for about 10% of the total mass of aerosols in our atmosphere. They have impacts on climate and precipitation that adversely affect human health (Vallero, 2014).

The particles are categorized according to their size. They can have any shape and can be solid particles or liquid droplets (Friedlander, 2000; Seinfeld and Pandis, 2016). The regulations and sampling methods focus on the size of the particles, since it is the main limiting factor for the greater or lesser penetration of particles in the respiratory tract. Therefore, the control networks carry out the determination of those particles less than 10 μm in diameter, called PM_{10}, which are those that have a greater capacity of access to the respiratory tract (respirable particles) and therefore affect it more. Within the PM_{10} fraction, the smallest particles (less than 2.5 μm in diameter, known as $PM_{2.5}$) are deposited in the alveoli, the deepest part of the respiratory system, being trapped and being able to generate more severe effects on health. In general, the thick part of the PM_{10} is largely made up of primary particles emitted directly into the atmosphere either by natural phenomena (forest fires or volcanic emissions) or by human activities (agricultural or construction work, dust re-suspension, industrial activities, etc.). On the other hand, fine particles ($PM_{2.5}$) are usually composed mainly of secondary particles formed in the atmosphere from a gaseous precursor (NO_x, SO_2, VOC, NH_3, etc.), by chemical processes or by reactions in liquid phase (Seinfeld and Pandis, 2016).

Over the last few decades, atmospheric particulate matter (PM) has become a relevant subject for research, due to its effects on human health and ecosystems. Nowadays it is well-known that particulate matter (PM) penetrates into the respiratory system, producing an increase in respiratory diseases and is responsible for harming lung and throat tissues (Turner et al., 2011). PM_{10} is defined as particulate matter having an effective aerodynamic diameter smaller than 10 μm, and is one of the most dangerous pollutants. Previous research had led to the conclusion that there is a correlation between PM_{10} levels and the increase in hospital admissions for lung and heart disease (Ostro et al., 1999). Nowadays there is clear scientific evidence (Dockery and Pope, 1994; Godish et al., 2014) to show that even small PM_{10} concentrations in ambient air can damage human health. These are the main reasons for the growing interest in the study of the life cycle of PM_{10} particles.

The European Commission has established health-based standards for PM_{10} and $PM_{2.5}$, as reported in the EU Air Quality Framework Directive (Directive 2008/50/EC, 2008), whereby the threshold for the daily average has been fixed at 50 μg/m^3, a figure not to be exceeded for more than 35 days in one year and with an annual upper limit of 40 μg/m^3 for PM_{10}.

One of the main sources of PM_{10} in Oviedo, as in most cities in the western world, is vehicular traffic, because of primary and secondary emissions from exhausts and suspended dust from the streets generated by circulation. Other sources are industrial activity and heating. Since these particles can penetrate the respiratory tract of humans due to their small size, they are potential agents of disease.

From our point of view, it is worth highlighting that a recent study (Ortiz et al., 2017) stated that in Spain, pollution killed on average 2683 people in recent years. The same study said that over the last decade 2963 people died in Oviedo due to these causes, which they link to concentration of PM_{10} and $PM_{2.5}$ particles. Mortality due to pollution in Oviedo is the third highest in Spain after San Sebastián and Madrid.

Finally, the organization of this paper is as follows: the dataset, materials and methods are described in Section 2; the discussion of the results is carried out in Section 3 and the conclusions appear in Section 4.

2. Materials and methods

2.1. Experimental dataset

The Government of the Principality of Asturias has a total of 19 ambient air monitoring stations. Three of them are located in the metropolitan area of Oviedo. The data analyzed correspond to the station called *Palacio de los Deportes*, (Sports Centre). This station was chosen as it corresponds to one of the most populated areas of the city where one of the motorways that comes into the city is located. In December 2015, the traffic on this motorway had to be restricted due to the number of days that the daily average of PM_{10} was over 50 μg/m^3. Please note that according to EU regulations the number of days with an average value over 50 μg/m^3 cannot exceed 35 days. Also, in 2016 and at the beginning of 2017, episodes of high PM_{10} concentration occurred in the metropolitan area of Oviedo and were reported by the local press. Fig. 1(a) shows the location of the three ambient air monitoring stations in the metropolitan area of Oviedo and a picture of the air monitoring station called *Palacio de los Deportes* (see Sports Centre in Fig. 1(b)).

Every 15 min, the ambient air monitoring stations record the concentrations of sulfur dioxide (SO_2), nitrogen monoxide (NO), nitrogen dioxide (NO_2), particulate matter with a diameter less than 10 μm (PM_{10}), ozone (O_3), bencene (C_6H_6), toluene (C_7H_8) and xylene or dimethylbenzene ($(CH_3)_2C_6H_4$) measured in μg/m^3, and carbon monoxide (CO) measured in mg/m^3 and so this information is available from the Air Quality Department of the Government of the Principality of Asturias. For the present research, information from 1st January 2010 to 31st July 2017 was retrieved. The information was processed using the average monthly values of the each of above pollutants supplied by the Air Quality Section of the Government of the Principality of Asturias as input data for the models.

2.2. Computational procedure

2.2.1. Support vector machine (SVM) method

Support vector machine (SVM) methods are supervised machine learning algorithms that can be used for regression and classification problems (Bishop, 2006; Cristianini and Shawe-Taylor, 2000; Hansen and Wang, 2005; Hastie et al., 2003; Li et al., 2008; Schölkopf et al., 2000; Steinwart and Christmann, 2008; Vapnik, 1998). If one is used as regressor, it is called *support vector regression* (SVR). Next, we want to estimate a value of the dependent variable y' that is typically real. The regression function $y = f(\mathbf{x})$ for a training dataset $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^{L}$ with $y_i \in \Re$ and $\mathbf{x}_i \in \Re^D$, so that L is the number of the samples in the training dataset and D is the dimension of the input dataset, can be expressed as:

$$f(\mathbf{x}_i) = \mathbf{w} \cdot \mathbf{x}_i + b \quad (1)$$

where $\mathbf{w} \cdot \mathbf{x}_i$ is the scalar product of the weight vector $\mathbf{w} \in \Re^D$ and $\mathbf{x}_i$, and b the intercept of the model indicating the bias. As a general rule, the SVM regression uses a so-called "penalty function" with value zero if the predicted value y_i is within a distance of less than ε from the observed value t_i, that is, if $|t_i - y_i| < \varepsilon$. The zone that satisfies this condition $y_i \pm \varepsilon$ for all i is the $\varepsilon-$ insensitive tube (see Fig. 2). Another modification to the penalty function is that the output variables falling out of the tube are supplied through two penalizations in the form of slack variables that depend on the position in relation to the tube: if they stay above (ξ^+) or below (ξ^-) where $\xi^+, \xi^- > 0$ for all i:

$$t_i \le y_i + \varepsilon + \xi^+ \quad (2)$$

$$t_i \ge y_i - \varepsilon - \xi^- \quad (3)$$

Therefore, the SVR error function is now expressed as follows (Cristianini and Shawe-Taylor, 2000; Hansen and Wang, 2005; Hastie