



Air pollution relevance analysis in the bay of Algeciras (Spain)

M. I. Rodríguez-García¹ · J. González-Enrique¹ · J. A. Moscoso-López² · J. J. Ruiz-Aguilar² · I. J. Turias¹

Received: 23 June 2021 / Revised: 11 March 2022 / Accepted: 5 August 2022
© The Author(s) 2022

Abstract

The aim of this work is to accomplish an in-depth analysis of the air pollution in the two main cities of the Bay of Algeciras (Spain). A large database of air pollutant concentrations and weather measurements were collected using a monitoring network installed throughout the region from the period of 2010–2015. The concentration parameters contain nitrogen dioxide (NO₂), sulphur dioxide (SO₂) and particulate matter (PM₁₀). The analysis was developed in two monitoring stations (Algeciras and La Línea). The higher average concentration values were obtained in Algeciras for NO₂ (28.850 µg/m³) and SO₂ (11.966 µg/m³), and in La Línea for PM₁₀ (30.745 µg/m³). The analysis shows patterns that coincide with human activity. One of the goals of this work is to develop a useful virtual sensor capable of achieving a more robust monitoring network, which can be used, for instance, in the case of missing data. By means of trends analysis, groups of equivalent stations were determined, implying that the values of one station could be substituted for those in the equivalent station in case of failure (e.g., SO₂ weekly trends in Algeciras and Los Barrios show equivalence). On the other hand, a calculation of relative risks was developed showing that relative humidity, wind speed and wind direction produce an increase in the risk of higher pollutant concentrations. Besides, obtained results showed that wind speed and wind direction are the most important variables in the distribution of particles. The results obtained may allow administrations or citizens to support decisions.

Keywords Air pollution · Regression analysis · Relevant variables · Relative risk · Spatial behavior

Introduction

The knowledge about how the globe is suffering from continuous atmospheric degradation is what motivates this study which is focused on air pollution in port-cities. Universally, several summits have been developed for decades to tackle shipping emissions. Since until 31st December 2019, for ships operating outside the ECA,¹ the limit for sulphides content of ships' fuel oil was 3.50% m/m (mass by mass); from 1st January 2020 onwards, 0.50% m/m limitation must be applied concerning the IMO² established in the 16th October 2008 committee. This deadline was set in the

MARPOL treaty. Mainly, air pollution caused by vessels and aviation are referred to sulphur dioxide (SO₂), nitrogen dioxide (NO₂) (Rivera et al. 2015) and particle matters (PM) (Chaloulakou et al. 2005; Agrawal et al. 2008; Grivas et al. 2018); thereby, this study is focused on these pollutants. Moored ships are also responsible for smoke in the air in port-cities, basically, due to the powerful engines usage in secondary electricity supply devices (Adamo et al. 2014). Moreover, European Directive 2008/50/EC (EU directive, 2008) establishes several thresholds and an AQI³ for every pollutant as mentioned in the study about particular matter with an aerodynamic diameter < 10 µm (PM₁₀) (Vicente et al. 2012).

As stated in (Westmoreland et al. 2007), nitric oxide (NO) together with NO₂ is known as NO_x and all kind of high-temperature combustions, as vehicles engines, are related to them mainly in urban areas (Chaloulakou et al. 2008). According to (Carslaw et al. 2007), NO₂ emissions are more associated with diesel engines. The extent of this research

Editorial responsibility: Samareh Mirkia.

✉ M. I. Rodríguez-García
inma.rodriiguezgarcia@gm.uca.es

¹ Department of Computer Science Engineering, PolytechnicSchoolofEngineering, University of Cádiz, Algeciras, Spain

² Department of Industrial and Civil Engineering, PolytechnicSchoolofEngineering, University of Cádiz, Algeciras, Spain

¹ ECA = Emission Control Areas.

² IMO = International Maritime Organization. <http://www.imo.org/>.

³ AQI = Air Quality Index.



is to assess how is the real air quality scope in a controversial zone to face future predictive studies (Turias et al. 2008; Munoz et al. 2014; González-Enrique et al. 2019a). To tackle researches about air quality, a complex scenario where several pollution sources interact was chosen. The Bay of Algeciras is a strategic point with several oil refineries, steel factories, and other industries. Algeciras port together with Gibraltar airport, which are connected by many roads with plenty of freight transport and constant private traffic, contributes to very complicated air pollution conditions. The second most populated city in the Bay is La Línea with 63,147 inhabitants in 2019. Algeciras is the major town in the Bay with a population of 121,957 inhabitants in 2019. Its port is of real importance not only in Spain but also in the world.

Overwhelming pieces of evidence show that particle pollution in the outdoor air we breathe, those coming from vehicles exhaust pipes (Crabbe et al. 1999; Carslaw et al. 2007; Bozkurt et al. 2018), coal-fired power plants, petroleum refineries and other industrial sources, can cause lung cancer and higher mortality rates in urban areas (Finkelstein et al. 2003). It was observed that several substances, including PM_{10} , reached higher levels in urban sites in the winter season (Bozkurt et al. 2018). Long-term pollution exposure to nitrogen oxides or sulphides can contribute to ailments such as cancer or asthma (Clench-Aas et al. 2000; Finkelstein et al. 2003). It is essential to control immissions since these affect human beings. Air pollutants are spread to different cities close to the emission points by winds (Cheung et al. 2020). Previous studies of SO_2 demonstrated that short-term were better than medium-term predictions, and the reverse in the case of PM_{10} concentrations (Turias et al. 2008). A recent study related to PM pollutants shows that even though when the heavy industries have decreased their manufacturing in the first lockdown period during the SARS-COV-2 (Covid-19) health crisis, severe pollution is not reduced when meteorology is adverse (Wang et al. 2020). Furthermore, it must be considered that if ozone (O_3) suffers from ozonolysis in presence of high levels of SO_2 and H_2O , the potential formation of secondary aerosols depends on relative humidity and meteorological conditions (Diaz-de Mera et al. 2017). Also, secondary aerosols (NO_2) are formed in the chemical reaction between the NO and O_3 (Westmoreland et al. 2007). Thus, air pollution in urban centres is a complex toxic-components mixture affected with the weather conditions and with a high impact on inhabitants, above all in those with cardiac insufficiency and respiratory distress (Kolehmainen et al. 2001). PM_{10} pollutant

enters the body exclusively through the respiratory system (Vicente et al. 2012). Therefore, it is primordial to identify the temporal evolution of pollutant concentrations in the air in urban regions to ensure the living standard. Immission data were collected from a monitoring network located in this study area by the Environmental Agency of Andalusian Government in the south of Spain. The study region has also a peculiar local meteorological scenario due to the closeness to The Strait of Gibraltar. The study contains descriptive statistical methods and more sophisticated statistical tools, such as *p-values* in regression analysis, or trend predictions.

Citizens or administrations have the need for reliable information about the possible risks they are exposed. One of the main aims of this study: knowing the most important variables and causes of high levels of pollution and, on the contrary, the variables that foster low levels of pollution. In this work, these challenges are achieved through a data-driven approach. Using historical data, a statistical analysis has been performed including three stages: i) descriptive, ii) predictive, and iii) prescriptive, which will be explained more deeply in the next sections.

The rest of this paper is organized as follows. Section 2 describes the data and the case study. Section 3 introduces the methodology. Section 4 presents and discusses the results, and finally, Sect. 5 establishes the main conclusions.

Materials and Methods

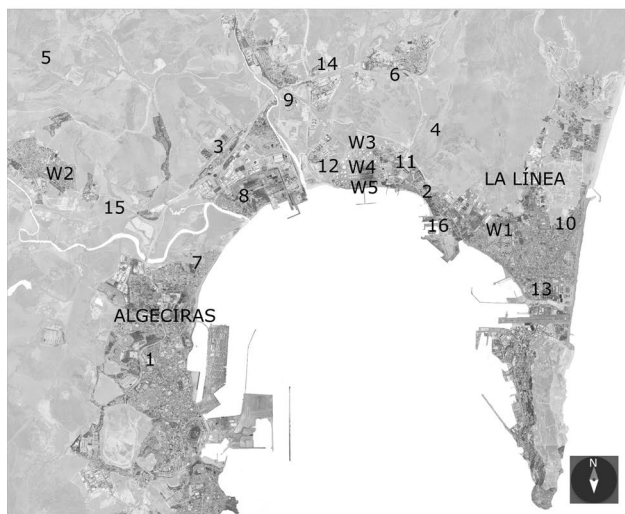
This section introduces a description of the database. A map of the area of study with reference to Spain is shown in Fig. 1(a), and a general location of the Bay of Algeciras is shown in Fig. 1(b). In this figure, the position of Algeciras port can be seen with its massive dimension, a total length of 17,750⁴ m amongst berths and seawalls. The situation of the two main cities of the bay, where this study focuses, Algeciras and La Línea, are located in front of each other as seen in Fig. 1(b). There are two dominant winds in this region, Levante (East) and Poniente (West), which seem to be produced by the situation of the bay in the proximity of the Strait of Gibraltar in connection with the Mediterranean Sea and the Atlantic Ocean. This special location of the bay creates a powerful air stream. Besides, as shown in Fig. 1(b), the pollutants and meteorological stations are spread over the bay.

Naturally, a higher amount of available resources in data collection, such as sampling points and sensors, will chase more reliable management of immission measuring.

⁴ <https://www.apba.es/datos-tecnicos>.



(a)



(b)

Fig. 1 Area of study with the monitoring stations: **a** Site location: Bay of Algeciras (South of Spain) and **b** Location of pollutants and meteorological monitoring stations. The monitoring stations codes are collected in Tables 1 and 2

Table 1 Meteorological monitoring stations

Code	Weather station description
W1	La Línea
W2	Los Barrios
W3	Cepsa 10 m high
W4	Cepsa 60 m high
W5	Cepsa 15 m high

The Andalusian Government maintains its air quality monitoring station system throughout the Bay, where pollutant

Table 2 Monitoring stations and the pollutants measured. NO₂ is recorded in a total of fourteen stations, SO₂ in sixteen, and PM₁₀ in ten stations

Code	Description of stations	Collected pollutants
1	Algeciras (EPSA)	NO ₂ , SO ₂ , PM ₁₀
2	Campamento	NO ₂ , SO ₂
3	Los Cortijos	NO ₂ , SO ₂ , PM ₁₀
4	Hostelería	NO ₂ , SO ₂
5	Alcornocales	SO ₂ , PM ₁₀
6	Carteya	NO ₂ , SO ₂ , PM ₁₀
7	Rinconcillo	NO ₂ , SO ₂ , PM ₁₀
8	Palmones	NO ₂ , SO ₂ , PM ₁₀
9	San Roque	NO ₂ , SO ₂ , PM ₁₀
10	El Zabal	NO ₂ , SO ₂ , PM ₁₀
11	Economato	NO ₂ , SO ₂
12	Guarranque	NO ₂ , SO ₂
13	La Línea	NO ₂ , SO ₂ , PM ₁₀
14	Madrevieja	NO ₂ , SO ₂
15	Los Barrios	NO ₂ , SO ₂ , PM ₁₀
16	Puente Mayorga	SO ₂

concentrations and atmospheric parameters were collected from the period of 2010 to 2015 and kindly provided to the University of Cádiz. The Bay of Algeciras counts on sixteen monitoring stations for collecting air pollutants and five meteorological specialized sensors. Three of the weather sensors are located inside a petroleum refinery (CEPSA) at three different heights (Table 1). Table 2 shows every pollutant in the monitoring stations. NO₂ pollutant is recorded in fourteen stations, SO₂ in sixteen stations, and PM₁₀ in ten stations.

The data have been collected hourly during the period of six years, from 2010 to 2015, with an apparent total database of 52,560 hourly data. Besides, twenty-four meteorological variables, described in Table 3, were collected hourly, as well. Following previous works (Turias et al. 2008; Munoz et al, 2014; González-Enrique et al. 2019b, 2019c), a procedure of missing data imputation was developed as a pre-processing step. The meteorological variables analysed are wind speed (WS), wind direction (WD), solar radiation (SR), atmospheric pressure (AP), rainfall (RF), relative humidity (RH), temperature (T) and the pollutants are NO₂, SO₂ and PM₁₀. These are the main substances expelled by the principal sources of air pollution in this area.

After data imputation, this study analyses the concentrations database together with the weather variables. Statistical parameters are obtained in every station to study their correlations and to study trend connections with other stations (descriptive approach). Then multivariate regression models

Table 3 Meteorological variables

Variable	Description
W1:WD	Wind direction (degrees)
W1:RH	Relative humidity (%)
W1:RF	Rainfall (l/m2)
W1:T ^a	Temperature (°C)
W1:WS	Wind speed (km/h)
W2:WD	Wind direction (degrees)
W2:RH	Relative humidity (%)
W2:RF	Rainfall (l/m2)
W2:AP	Atmospheric pressure (hPa)
W2:SR	Solar radiation (w/m2)
W3:RH	Relative humidity (%)
W3:RF	Rainfall (l/m2)
W3:AP	Atmospheric pressure (hPa)
W3:SR	Solar radiation (w/m2)
W4:WD	Wind direction (degrees)
W4:T ^a	Temperature (°C)
W4:WS	Wind speed (km/h)
W5:WD	Wind direction (degrees)
W5:WS	Wind speed (km/h)
W5:RH	Relative humidity (%)
W5:RF	Rainfall (l/m2)
W5:AP	Atmospheric pressure (hPa)
W5:SR	Solar radiation (w/m2)
W5:WS	Wind speed (km/h)

Table 4 Descriptive statistical measurements in Algeciras (1) and La Línea (13) monitoring stations

Station	Mean ($\mu\text{g}/\text{m}^3$)	Median ($\mu\text{g}/\text{m}^3$)	σ^2	Kurtosis	Skewness
<i>NO₂</i>					
1	28.850	25.000	460.680	3.587	0.910
13	26.189	19.833	408.510	4.669	1.331
<i>SO₂</i>					
1	11.000	8.333	67.818	39.726	3.867
13	11.966	10.000	64.693	19.789	2.813
<i>PM₁₀</i>					
1	27.409	24.543	295.620	25.456	2.803
13	30.745	27.333	405.500	67.438	5.140

(Romero et al. 2020) have been established for every pollutant in each station to check the relevant features and to dispose of estimation models to choose a virtual sensor with numerous applications, such as, missing data imputation,

a real-time usage in a robust monitoring net or prediction of the air pollution (a predictive approach). Previously, authors have used these methods (Turias et al. 2008; Munoz et al. 2014; Moscoso-López et al. 2019; Ruiz-Aguilar et al. 2021) in different works. Finally, measuring higher values of immission pollutants concentrations, a complete assessment of relative risks was developed to have at our disposal cause-effect knowledge about which are the main hazardous variables to prevent and take decisions (a prescriptive approach).

Descriptive analysis

Correlation

An in-depth statistical assessment was developed to get a general idea of the scope of the pollution in the study area. In Table 4, the mean, median, mode, variance, kurtosis and skewness are collected for every pollutant and study city. On the other hand, a linear correlation analysis was developed. A correlation is a reciprocal relation between two or amongst different variables which are expected to have some kind of connection, even though correlation does not mean relation. For instance, if one of these variables grows, the other is expected to increase or even decrease. Multidimensional correlation results were computed (see Fig. 2a, b and c). Pearson correlation coefficient (r) is shown in Eq. (1) where the numerator corresponds to $COV(X, Y)$, the covariance between every pair of independent variables, and the denominators are, respectively, $\sigma_x \cdot \sigma_y$, the typical deviations of them.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_i)^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y}_i)^2}} \quad (1)$$

Mutual information

Mutual information (MI) (Shannon, 1948) is based on Shannon's information theory. MI measures the statistical dependence between two variables, and thus, independency implies a low mutual information between them. Mutual information is a nonnegative measure (Kullback, 1968). MI uses the concept of entropy as a measure of uncertainty, since its maximum is when all values have equal probability of occurrence (Shannon, 1948). The advantage of mutual information is its ability to estimate a general dependence

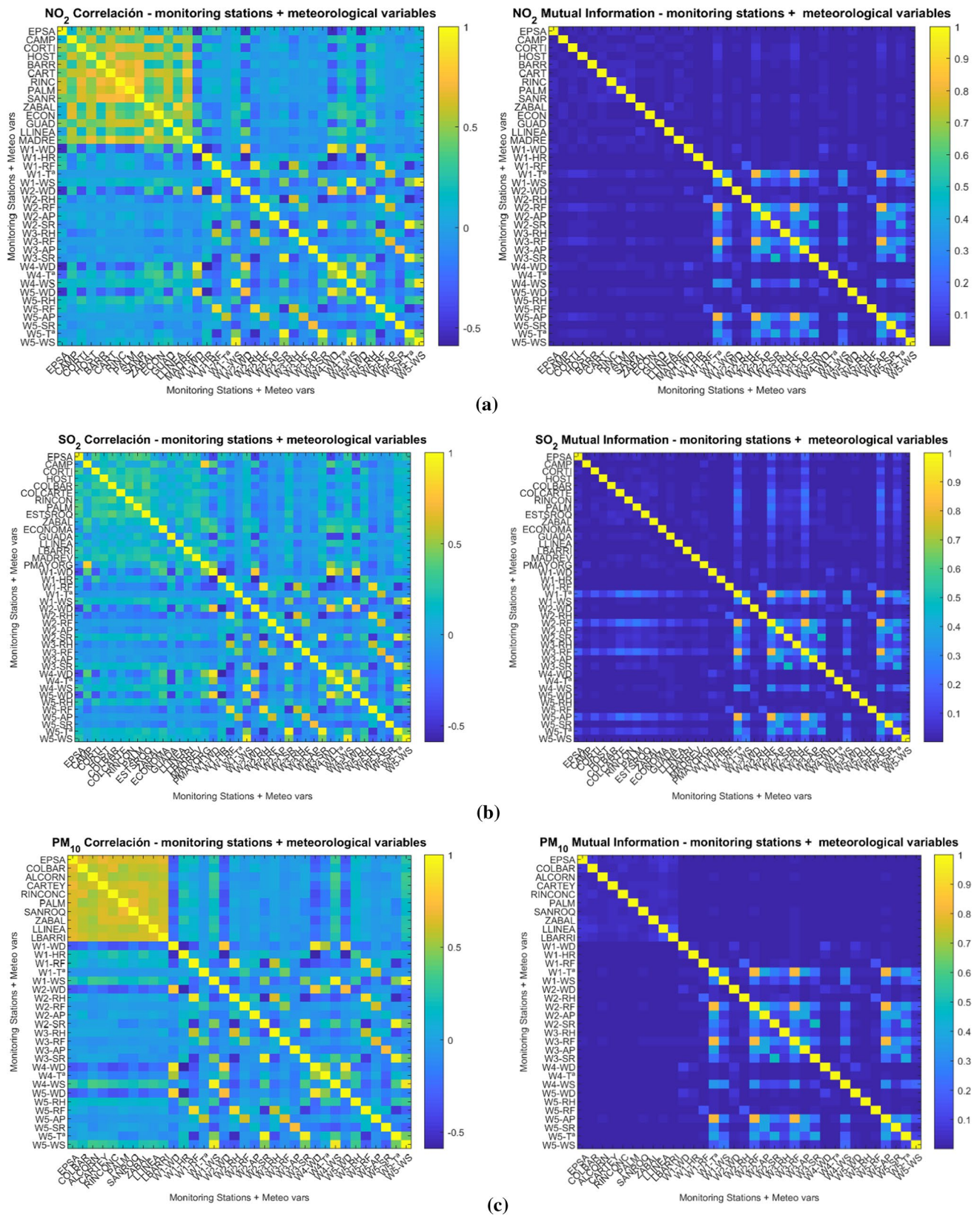


Fig. 2 Plots of correlation values and mutual information among pollutants monitoring stations and meteorological variables **a** NO₂, **b** SO₂, **c** PM₁₀

Table 5 Trend tests for similarity amongst monitoring stations

<i>Daily data test</i>			
Test	NO ₂	SO ₂	PM ₁₀
Wilcoxon rank	1,13	1,13,15	1,10,15; 7,8,13
Wilcoxon signed	–	1,13,15	1,15; 7,10,13
Wilcoxon signed-rank	–	1, 15	7,10,13
T test	–	1, 15	1,15; 7,8,10,13
<i>Weekly data test</i>			
Wilcoxon Rank	–	1, 15	7, 10, 13
Wilcoxon Sing	–	1, 15	–
Wilcoxon signed-rank	–	1, 15	10,13
T test	–	1, 15	7,10,13

between variables, as opposed to correlation between variables where only linear relationships are considered.

To specify how much common certainty there are in two data samples, MI uses the concept of crossentropy $H(X, Y)$:

$$MI(X, Y) = H(X) + H(Y) - H(X, Y) \quad (2)$$

where $H(X)$ is the entropy of x variables and $H(Y)$ is the entropy of the output y (Cover and Thomas, 1993).

MI can be computed using Eq. 2 through an estimation adaptive partitioning (Darbellay and Vajda, 1999) procedure as authors previously used in González-Enrique et al. (2021). The ITE Toolbox (Szabó, 2014) has been applied in this work for MI calculation.

Trend analysis

In order to study trends between the monitoring stations for each air pollutant, we have used nonparametric trials and non-normality tests such as Wilcoxon rank test, Wilcoxon signed test and Wilcoxon signed-rank test to be compared with T-test (Box et al. 1976). Data are grouped into pairs of observations or monitoring stations, (x_i, y_i) , which are aimed at getting to know if every pair is equal or not from a statistic point of view. The assumptions in these models are that if $y_i - x_i \neq 0$, samples are independent, and they have a continuous and symmetric distribution concerning the same common median θ . These tests consist in returning a logical value indicating the test decision within a confidence interval, conventionally, of 95%. The decision will vary according to the kind of test we use. Basically, if the logical value it returns is 1, this indicates a rejection of the null hypothesis, and if the logical value it returns is 0, this indicates a failure to reject the null hypothesis at 5% of the significance level. *Wilcoxon rank test* is used when two samples have different lengths and they are independent. It tests the null hypothesis that data are samples

from continuous distributions with equal medians, against the alternative that they are not. *Wilcoxon sign test* is the simplest nonparametric test applied in paired samples (Wasserman, L., 2004a). It is usually used to test the median of a population and returns the *p-value* for a two-sided sign test through a binomial distribution (Wasserman, L., 2004b). *Wilcoxon signed-rank test* is used instead of a T-test when normality in the sample cannot be proved, besides this test is the nonparametric of the dependent samples *T test*. A T test is used to compare two groups determining if there is a significant difference between their means, which might be related to certain features. Results of the several tests applied are collected in Table 5.

Predictive analysis

Linear regression is a linear approach to modelling the relationship between two variables, used as a predictive tool for studying the response of one dependent variable, or response Y , and the covariate variable X , also called feature or predictor (Eq. 3).

$$R(x) = E(Y/X = x) = \int yf(y/x)dy \quad (3)$$

Multiple regression is a *many-to-one* relationship amongst independent variables x_i and the response y . Adding more predictors to the model does not mean a better response, and it may produce *overfitting* or *multicollinearity*. The sum of linear parameters and an error gives the multiple regression model (4).

$$y_i = \sum_{i=1}^n (\beta_0 + \beta_i \cdot X_{ij}) + \epsilon_i \quad (4)$$

where β_0 is the intercept, β_i are coefficients which show the weight of every independent variable X_i , and ϵ_i ⁵ is the error or residual. When this equation is estimated, we get the predicted value of the dependent variable, \hat{y}_i , obtaining $b_0, b_1, b_2, \dots, b_i$, the estimates of $\beta_0, \beta_1, \beta_2, \dots, \beta_i$, respectively.

As it is well known, the coefficient of determination, R^2 , is preferably closer to one as possible. In addition, the *p-value* of each variable was calculated in order to discard all of the redundant independent variables whose *p-value* is ≥ 0.01 . Generalization is key in data analysis since it is the ability to get good results with new unseen data. Building a model that fits the data sorely well does not always guarantee that the model is useful. The model should work well not only with the data it has learned from but also with the data that have not been used so far. The procedure to measure generalization is to divide our data into two sets: the training

⁵ ϵ_i is the difference between the observed and the predicted values.

set and the test set. The training set, larger than the test set, contains the data used by the model to learn the parameters, and the test set contains unseen data not used in learning. The test set will be used to know how the model behaves with new data. In this analysis, the period 2010–2014 is used as a training set and the year 2015 is used as the test set. The results have been collected to these “new” data.

Prescriptive analysis

An analysis can be descriptive or predictive when representing an event, or prescriptive when researching a cause-effect occurrence, and this fact can be used in order to have a support decision tool (Schmidt and Kohlmann, 2008). The consequence can be evaluated by the relative risk analysis, which is explained briefly here. There are two methods for measuring the risk: *Odds Ratio* and *Risk Ratio*, both dimensionless and accompanied by the confidence interval (CI), which is a measure of the precision of the estimation (Tripepi et al. 2007). Their usage depends on the design of our study.

Odds Ratio (OR) it is used when a retrospective design is applied. This means that the first step is focusing on the consequences and then analysing the causes. The odds are a way of representing probability (Tripepi et al. 2007). OR is also known as “Cases and Controls” (Ganguly 2006). Basically, odds is the ratio that represents the probability of occurrence of an event by means of the quotient between the happening event probability and the non-happening event probability (Bland and Altman 2000). It indicates how higher is the probability of occurrence of an event towards its non-occurrence (Sumargo 2018). These terms are described below.

$$OR = \frac{\text{Odds}_1}{\text{Odds}_0} = \frac{R_1/(1 - R_1)}{R_0/(1 - R_0)} = \frac{a_1 b_0}{a_0 b_1} \quad (5)$$

Risk Ratio (RR) also called relative risk. It is a statistical concept used as a measure of association between dependent and independent variables. It is indicated to prospective studies, beginning in the reviews of the causes and their supervision until examining the consequences. RR is also called “Cohort study” (Finkelstein et al. 2003; Schechtman 2002). The relative risk can be calculated as the ratio between the two incidence proportions or two incidence rates (Tripepi et al. 2007). It is the quotient between the proportion of cases with risk factor (subindex 1) and the proportion of cases without risk factor (subindex 0).

$$RR = \frac{R_1}{R_0} = \frac{a_1/n_1}{a_0/n_0} \quad (6)$$

The terms are calculated counting and separating the individuals of the sample in cases-no factor, cases-factor, no cases-no factor, and no cases-factor. The total of individuals with no factor is n_0 and the total of individuals with factor is n_1 . The sum of individuals with cases-no factor is a_0 , the sum of individuals with cases-factor is a_1 , the sum of individuals with no cases-no factor is b_0 , and the sum of individuals with no cases-factor is b_1 . Thus, it is easy to imagine that $n_0 = a_0 + b_0$ and $n_1 = a_1 + b_1$.

For instance, the relative risk of a situation is the ratio of risks of the treated group and the control group (Schechtman 2002). Observing Eqs. (5) and (6) is immediate to realise that when $OR > 1$ or $RR > 1$, a positive association, the presence of the factor is related to a higher occurrence of the event, and the reverse if the association is negative, $OR < 1$ or $RR < 1$ (Finkelstein et al. 2003). Conversely, when $OR = 1$ or $RR = 1$, there is no association between the presence of the factor and the event. There are several existing linkages between OR and RR. RR is more perceptive. OR lets us adjust by confounding variables through logistic regression, although it is not applied in this study.

Results and discussion

An overview of the in-depth analysis is shown in this section. Firstly, a descriptive statistical assessment was performed showing the most relevant features. Likewise, a linear correlation analysis was conducted to find out and determine which pairs of variables shared information. Also, a trend analysis was developed in order to establish those stations with similarities. Secondly, a linear multiple regression analysis was performed allowing to determine which variables are the most explanatory for every pollutant and location (with $p\text{-value} < 0.01$), and besides, to be able to establish how much variability can be fully explained with these models. These regression models can serve as virtual sensors because they are capable of inferring a measure (the concentration of a pollutant in a certain location) based on other variables. Therefore, they can be used for the imputation of missing data or as robust control in a monitoring network. Similarly, or with the same objectives, an approximation has been made to calculate which monitoring stations are equivalent to others, in this case using statistical comparison tests on their means.



UNECE⁶ announced that air pollution is now considered to be the world's largest environmental health threat, accounting for 7 million deaths around the world every year. The main substances affecting health are: nitrogen oxides (NO_x), sulphur oxides (SO_x), ozone and particulate matter. Both extent and duration of the exposure influence health diseases. This study and its results are interesting in order to give information about air pollution to citizens. Besides, regarding the economic costs of air pollution, WHO⁷ and OECD⁸ estimated in 2015 that the amount of money related to premature deaths and disabilities in Europe reached almost USD 1.6 trillion. Therefore, preventing long exposures could be useful to avoid risk factors for major diseases.

Finally, some results are collected on the statistical risks of interactions of the variables. Thus, we have obtained the most influenced interaction values for pollutants (especially in the case of the highest pollution values or, in contrast, of the lowest).

Descriptive analysis

The statistical terms of mean, median, variance, kurtosis, and skewness are collected in Table 4 for each pollutant recorded in the stations in both cities. Simply looking at this board we can have an idea of their relevance. Generally, the means of pollutant concentrations are very similar in Algeciras (number 1) and La Línea stations (number 13). Roughly, during the study period, it was observed that the highest mean occurs in La Línea for all pollutants except for NO₂. All medians are beneath the mean concentrations which indicate that fifty percent of concentration values are upper the median, proof that the database does not follow a normal distribution. This is also appreciated with a positive skewness (> 0) when symmetry tends to values higher than the mean. The sharpest graph is the one with the highest kurtosis coefficient, corresponding to the PM₁₀ in La Línea monitoring station (13).

Correlation

Figure 2 exposes the correlation values for every pair of variables. These plots represent the resulting correlation results, showing in colours closer to yellow the values of highest correlation coefficients tending to one. Coefficients of one, those of bright yellow in the diagonal, correspond to every variable with itself (it is not representative). Hereafter, the most significant correlation values in Algeciras and La Línea stations with the rest monitoring stations and

atmospheric variables are exposed here. Speaking in absolute values, the correlations do not exceed $r=0.6781$, which correspond to Carteya station correlated with Algeciras station for the PM₁₀ pollutant and, continuedly, also the stations Los Barrios and El Zabal show similar correlation. For this pollutant, similar values of correlations are depicted in La Línea for Los Barrios, Algeciras, and El Zabal stations. Observing weather variables for PM₁₀, wind direction is the highest correlated variable in Algeciras station and wind speed in La Línea. Moreover, regarding weather variables, the highest values of correlation connect wind direction in weather stations W1, W2, and W4 with Algeciras station for NO₂. Besides, this pollutant in Algeciras station presents a correlation with Palmones, El Rinconcillo, and Los Barrios stations. On the other hand, La Línea station presents a connection with Campamento, Escuela de Hostelería, and Los Barrios stations for the same pollutant. In the case of La Línea station for the NO₂ pollutant, the weather variables more correlated are wind direction in W1, W2 stations, and temperature in W4. Finally, for SO₂ wind direction is the weather variable that most correlates both cities and also relative humidity in La Línea station measured in W3. Considering monitoring stations for SO₂, Algeciras is correlated with Palmones, Los Cortijillos, and San Roque stations. La Línea is correlated with El Zabal, Puente Mayorga, and Campamento stations, which makes sense due to their proximity.

Mutual information

In order to get knowledge about the nonlinear behaviour amongst the variables, MI has been computed. In this sense, MI provides a different and more general criterion for investigating relationships between variables.

Regarding to the MI results between the studied variables, it is observed that there is not much nonlinear information between the monitoring stations. The maxima were found in the combinations of the meteorological variables W1-T, W2-RF, W3-RF, and W5-AP. This information complements to that calculated with the linear correlation and allows us to assume that these variables could be used in nonlinear regression or prediction models. Nevertheless, most of the shared information is linear rather than nonlinear with the limits of values observed in Fig. 2 (maximum 0.7 approximately). The behaviour is very similar for the three studied pollutants, with higher MI values only in the case of SO₂ in the monitoring stations when a combination with the variables W1-T, W2-RF, W3-RF, and W5-AP was tested.

⁶ United Nations Economic Commission for Europe.

⁷ World Health Organization.

⁸ Organisation for Economic Co-operation and Development.



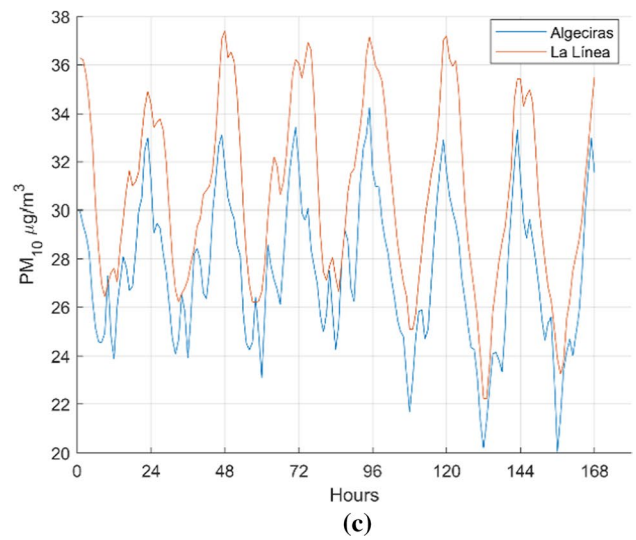
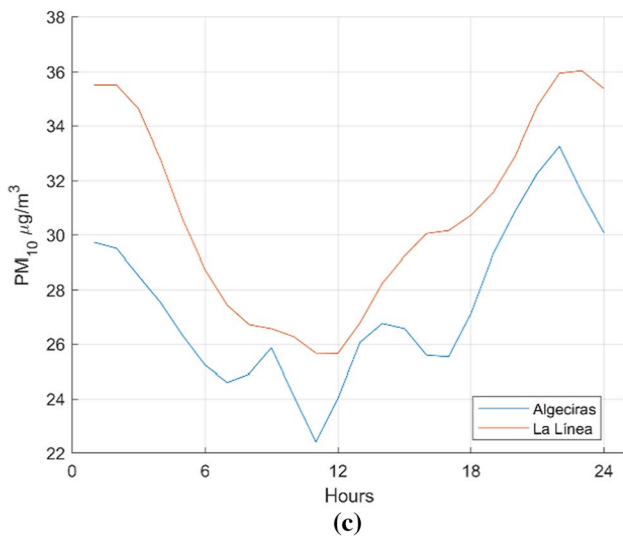
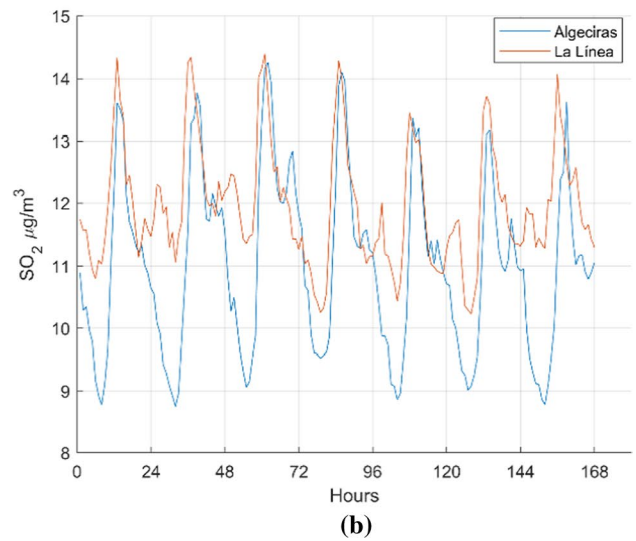
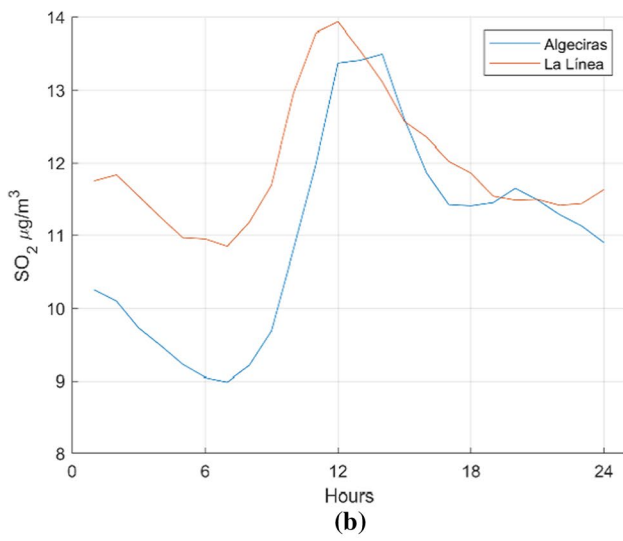
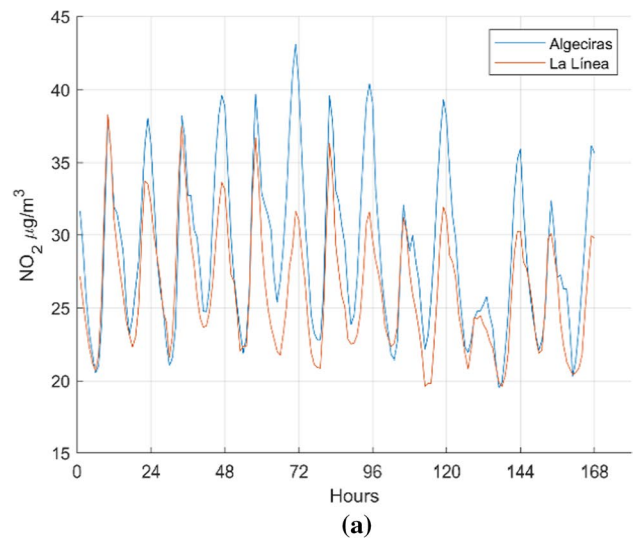
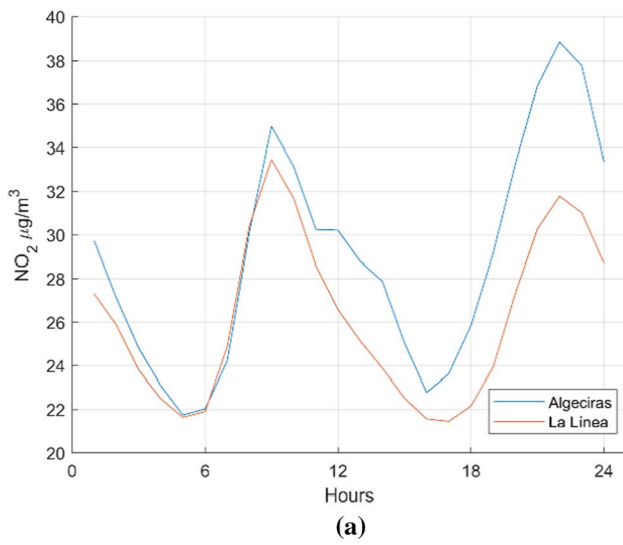


Fig. 3 Average day. Comparison between the hourly mean concentration of pollutants in Algeciras and La Línea during the period 2010–2015: **a** NO₂ **b** SO₂ **c** PM₁₀

Fig. 4 Average week. Comparison between the hourly mean concentrations of pollutants in Algeciras and La Línea during the period 2010–2015: **a** NO₂ **b** SO₂ **c** PM₁₀

Table 6 Multiple linear regression values for every pollutant in Algeciras (1) and La Línea (13) monitoring stations

Pollutants	Monitoring stations	R ² (MLR)	MSE(MLR)
NO ₂	1	0.7360	269.14
	13	0.8655	150.60
SO ₂	1	0.6382	33.82
	13	0.5346	30.03
PM ₁₀	1	0.8244	96.92
	13	0.6608	221.09

Trend analysis

The tendency is a measurement of variability between two samples regarding to their means. In this section, two samples coming from different stations are measured in order to obtain if they have the same trend or not from a statistical point of view, using, henceforth, hypotheses test for both daily and weekly means. In daily calculation, all database is divided into twenty-four intervals, or hours, and the mean is calculated for every hour (see Fig. 3). In weekly calculation, data are divided into groups of 52 weeks in order to calculate the mean of every interval that is the weekly mean for the total of the six years (see Fig. 4). The figures show daily data that give the hourly mean concentration for every pollutant in Algeciras (EPSA) and La Línea (Fig. 3), and weekly data giving the daily mean concentration (Fig. 4). In these figures, we can see that the highest levels of SO₂ and PM₁₀ pollutants mainly affect the city of La Línea, and for the NO₂ pollutant, the highest level is obtained in Algeciras. Checking Fig. 3(a), a NO₂ concentration level peaks appears at 38.88 µg/m³ which corresponds to the city of Algeciras in contrast to the highest concentration level peak of 31.79 µg/m³ in La Línea. In both cities, the NO₂ concentrations grow strikingly at 10 p.m., together with a lower peak at 10 a.m. In Fig. 4(a) is noticed that the highest level of NO₂, 43.17 µg/m³, corresponds to Algeciras achieved on Wednesday, and in La Línea the highest level of NO₂ is 38.29 µg/m³ obtained on Monday. In the case of SO₂, Fig. 3(b) shows in La Línea a steadily rise until the highest value of 13.94 µg/m³ at 12 a.m. and very close, 13.49 µg/m³ at 2 p.m. in Algeciras. Figure 4(b) shows that Wednesday is the day of highest values, very similar both in La Línea (14.39 µg/m³) and Algeciras (14.26 µg/m³). Concerning PM₁₀ pollutant, Fig. 3(c) depicts a steep downward in both cities at 12 p.m. and also the highest mean-hour is 36.04 µg/m³ corresponding to La Línea at 11 p.m. In Algeciras, the value is 33.28 µg/m³ at 10 p.m.

Figure 4(c) shows the highest value of 37.41 µg/m³ in La Línea corresponds to Tuesday and in Algeciras 34.27 µg/m³ the Thursday. In general, at Algeciras station exists a major fluctuation for every pollutant, obtaining the lowest values of concentrations. At La Línea station, the values are maintained higher, probably caused by its proximity with industrial environment combined to west winds. On the other hand, after applying several trend-tests for both daily and weekly data to get those stations with similar behaviour, promising trend results were obtained for NO₂, SO₂ and PM₁₀ in the cities of Algeciras and La Línea (see Table 5). Trend results for similar stations demonstrate that for SO₂ pollutant Algeciras station (number 1) is always similar to Los Barrios station and in the case of Wilcoxon rank and Sign tests is also equal to La Línea station (number 13). This argument is used in the rest of tests. Regarding PM₁₀, the situation is strongly different since several stations show similar behaviour with all tests except Wilcoxon sign for weekly data test, which presents none of them.

Predictive analysis

A Multiple linear regression (MLR) analysis was applied in order to estimate the three different air pollutants as a function of meteorological variables and the rest of monitoring stations. In this work, the estimation of the parameters was done using the data in the period 2010–2014 (training or design set) and the regression results were collected using the year 2015 (as a test set). The best regression model is obtained when highest R² and lowest MSE. Regression results in Algeciras and La Línea are displayed in Table 6. We observe that in Algeciras station the best regression value is for PM₁₀ and in La Línea for NO₂, nevertheless, SO₂ presents similar values in both cities. Finally, using the *p-value* < 0.01 of the regression, the most relevant features can be selected in every model. Their weights or estimates (β_i) and the intercept (β_0) of each regression equation are shown in Table 7 along with the most relevant variables for every pollutant and station. Those variables with positive estimates indicate that the dependent variable (i.e. each air pollutant concentration) are positively affected by these variables and reversely with negative sign. For instance, regarding the weather variables, Algeciras station is positively influenced by wind speed in W4 and negatively in W1 for all pollutants. Considering the regression among monitoring stations, Algeciras is more positively affected by Los Barrios for all pollutants and also by Palmones station in the case of SO₂. Besides, Algeciras is affected negatively by San Roque for NO₂. In the case of La Línea station, the variables that



more influence positively are wind speed (in W1 and W4 weather variables) for all pollutants and also wind direction (in W1 for SO₂). Regarding monitoring stations, La Línea is positively affected by El Zabal in the case of NO₂ and SO₂ pollutants and Los Barrios station in the case of PM₁₀.

In contrast, La Línea station is negatively affected by Cortijillos station for the NO₂ pollutant. Negatively, the weather station that more leverages La Línea station is W5, wind speed (in the case of NO₂ and PM₁₀) and wind direction in the case of SO₂. The rest of the values can be depicted in the same way. Tables 1, 2, 3, and Fig. 1 can be checked to locate every station and to observe these interesting relations.

Prescriptive analysis

In this study, the relative risks RR > 1.25 have been considered as long as statistically significant with a *p-value* < 0.05 in a *Chi test*. The highest and significant RR values in stations are shown in Table 8. Considering the NO₂ pollutant measured in W2, lower degrees (first quartile, Q₁) of wind direction (WD) produce the highest relative risk (3.08) of suffering from an elevation of this pollutant above the mean

Table 8 Highest relative risks computed between pollutants and meteorological variables. Risk 0 means non-risk (protection), risk 1 means risk

Pollutants	Monitoring stations	Meteorological variables	Risk	Quartile	RR
NO ₂	1	W2:WD	1	1	3.08
	13	W2:WD	0	3	2.96
	13	W1:WD	0	3	2.35
SO ₂	1	W1:RH	1	4	3.48
	1	W5:WS	1	4	3.30
	13	W4:T	0	4	2.60
PM ₁₀	13	W4:WD	0	4	2.58
	1	W5:WS	0	4	2.65
	1	W4:WS	0	4	2.63
	13	W4:WS	1	4	2.31
	13	W1:WS	1	4	2.31

in Algeciras station. Furthermore, we observe that higher degree values of wind direction (third quartile, Q₃) in W2 do not present any risk of enduring a rise of NO₂ in La Línea station above the mean. In fact, they act as a protection in

Table 7 Relevant weather variables and monitoring stations (with regression *p-values* < 0.01) are shown for each pollutant in monitoring stations of Algeciras (1) and La Línea (13). β_{*i*} are the estimates, and β₀ is the intercept of the regression models

Pollutants	Stations	Relevant weather variables	β _{<i>i</i>}	Relevant stations	β _{<i>i</i>}	β ₀	
NO ₂	1	W4:WS	0.6363	15	0.3207	-146.6	
		W5:WS	0.2200	7	0.3225		
		W1:RF	0.2151	9	-0.2004		
		W1:T	-0.2429	14	0.1564		
	13	W1:WS	-0.4134	10	0.1448		
		W4:WS	0.3890	10	0.8554		-21.678
		W1:RH	0.3036	4	0.0915		
SO ₂	1	W1:RH	-0.3871	14	0.0827	-72.443	
		W5:WS	-0.3730	3	-0.0605		
		W4:WS	0.4007	8	0.1486		
	13	W1:RF	0.0956	15	0.1022		
		W1:WS	-0.1457	7	0.0938		
		W1:WD	0.1462	10	0.1317		-64.369
		W4:WS	0.1009	15	0.1309		
PM ₁₀	1	W1:RH	-0.1441	1	0.1112	-76.698	
		W5:WD	-0.1713	3	0.0378		
		W4:WS	0.6752	15	0.1858		
	13	W2:RH	0.1019	6	0.1530		
		W1:T	-0.1658	5	0.1241		
		W1:WS	-0.4414	13	0.0975		
		W1:WS	0.9279	15	0.2586		-25.918
		W1:T	0.2768	1	0.1720		
		W2:RF	-0.3222	5	0.1489		
		W5:WS	-0.3846	7	0.1348		

this station for this pollutant and weather variable and also occur to WD recorded in W1. Focusing on the SO₂ pollutant, if the relative humidity (RH) is high, measured in weather station W1 (this means that it is located in fourth quartile, Q₄), produces the existence of high relative risk (3.48) to suffer from SO₂ rise (having a value greater than the mean) in Algeciras station as also happens to wind speed measured in W5 station. Nevertheless, higher values of temperature and wind direction (Q₄) recorded in W4 station present protection in La Línea station for SO₂. In the case of PM₁₀, in Algeciras station higher values (Q₄) of wind speed measured in W5 and W4 stations act as a protection of not suffering from an elevation above the mean concentration. However, in La Línea station higher values (Q₄) of wind speed, in W4 and W1, do affect the risk increase of undergoing overruns above the mean.

Conclusion

An exhaustive statistical analysis in a complex scenario characterized by a real industrial region along with a quite singular meteorological situation was performed in this study. In general terms, in the period analysed it is observed that Algeciras station recorded higher values of NO₂ pollutant than La Línea station, however, considering SO₂, the values were very similar. In the case of PM₁₀, La Línea station collected the highest values.

Regarding regression models, the hypothesis that winds are important in this area is tested in this study, showing the relevance of correlation and highest regression coefficients for the two wind components (speed and direction) measured in different weather stations amongst other variables. Regression rates go from values of R² above 0.5346 to the max R² of 0.8655 that corresponds to NO₂ with La Línea station. Moreover, wind speed appears to be the most relevant variable in the majority of cases in both cities. The two wind components trigger an apparent particle movement which leads to an air-cleaning effect in The Bay of Algeciras. However, Algeciras is highly affected by east winds (Levante) and La Línea is more impacted by west winds (Poniente). Generally, considering the monitoring stations, Los Barrios station presents higher affection with Algeciras for all pollutants. In La Línea, El Zabal station seems to be more relevant. Both of them could be used with relevance in a robust virtual sensor of Algeciras and La Línea stations along with wind variables.

Considering relative risk results, Algeciras is affected negatively by lower degrees of wind direction for the NO₂

pollutant. For SO₂ the variables that affect the most to Algeciras are higher values of relative humidity, which produce a rise of SO₂ above the mean. This increase might be in concordance with the research of (Díaz-de Mera, 2017). In addition, higher values of wind speed produces also a relative risk in Algeciras. Nevertheless, in the case of PM₁₀, the highest values of wind speed cause protection leverage in this city, just the reverse than in La Línea station, where higher values of wind direction and higher temperatures act as protection. Admittedly, PM₁₀ pollutant is spread from Algeciras to La Línea through wind speed.

According to trend analysis, NO₂ pollutant presents similarities between Algeciras and La Línea stations. In the case of SO₂, the similarity is between Algeciras and Los Barrios station and with regard to PM₁₀ it can be measured indistinctly in Algeciras-Los Barrios, El Rinconcillo-La Línea or El Zabal-La Línea stations. These results, in general, are confirmed with the regression models.

According to the OECD, outdoor air pollution could cause 6–9 million premature deaths a year by 2060 and cost 1% of global GDP⁹—around USD 2.6 trillion annually—as a result of sick days, medical bills and other issues.

In accordance with recent estimates by the WHO, exposure to air pollution is thus a more important risk factor for major diseases. New tools such the one presented in this work have proven to be an effective tool in avoiding hazardous situations. Preventing potentially health risk events helps citizens in preventing morbidity and premature mortality, one of the targets under Sustainable Development Goal (objective 3) on good health and promoting wellbeing.

To summarise the above, an air pollution modelling approach based on different perspectives (descriptive, predictive, and prescriptive) was performed on this singular area using different statistical methods. The results obtained can be used as a virtual sensor in the case of sensor failures and also as a support decision tool for institutions and citizens to prevent peak-situations. Furthermore, this proposed approach could be used in different regions or scenarios in future researches.

Acknowledgements This work is part of the research project RTI2018-098160-B-I00 supported by 'Ministerio de Ciencia e Innovación (MICINN). Programa Estatal de I+D+i Orientada a Los Retos de la Sociedad'. Data used in this work have been kindly provided by the Andalusian Regional Government.

Declarations

Conflict of interest The authors have not got relevant conflicts of interest to declare to the content of this article.

⁹ Gross Domestic Product.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adamo F, Andria G, Cavone G, De Capua C, Lanzolla A, Morello R, Spadavecchia M (2014) Estimation of ship emissions in the port of taranto. *Meas J Int Meas Confed* 47(1):982–988
- Agrawal H, Welch W, Miller J, Cocker D (2008) Emission measurements from a crude oil tanker at sea. *Environ Sci Technol* 42(19):7098–7103
- Bland JM, Altman DG (2000) The odds ratio. *BMJ* 320(7247):1468
- Box GEP (1976) *Time series analysis: forecasting and control*, 5th edn. Wiley, New Jersey
- Bozkurt Z, Gaga EO, Taşpınar F, Arı A, Pekey B, Pekey H, Doğeroğlu T, Özden Uzmez O (2018) Atmospheric ambient trace element concentrations of PM10 at urban and sub-urban sites: source apportionment and health risk estimation. *Environ Monitor Assessment* 190(3):168
- Carslaw DC, Beevers SD, Tate JE (2007) Modelling and assessing trends in traffic-related emissions using a generalised additive modelling approach. *Atmos Environ* 41(26):5289–5299
- Chaloulakou A, Kassomenos P, Grivas G, Spyrellis N (2005) Particulate matter and black smoke concentration levels in central Athens Greece. *Environ Int* 31(5):651–659
- Chaloulakou A, Mavroidis I, Gavriil I (2008) Compliance with the annual NO2 air quality standard in Athens. required NOx levels and expected health implications. *Atmos Environ* 42(3):454–465
- Cheung CW, He G, Pan Y (2020) Mitigating the air pollution effect? The remarkable decline in the pollution-mortality relationship in Hong Kong. *J Environ Econ Manag* 101:102316
- Clench-Aas J, Bartonova A, Klaboe R, Kolbenstvedt M (2000) Oslo traffic study – part 2: quantifying effects of traffic measures using individual exposure modeling. *Atmos Environ* 34:4737–4744
- Crabbe H, Beaumont R, Norton D (1999) Local air quality management: a practical approach to air quality assessment and emissions audit. *Sci Total Environ* 235(1):383–385
- Diaz-de Mera Y, Aranda A, Martinez E, Rodriguez AA, Rodriguez D, Rodriguez A (2017) Formation of secondary aerosols from the ozonolysis of styrene: effect of SO2 and H2O. *Atmos Environ* 171:25–31
- EU (2008) Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe
- Finkelstein MM, Jerrett M, DeLuca P, Finkelstein N, Verma DK, Chapman K, Sears MR (2003) Relation between income, air pollution and mortality: a cohort study. *CMAJ* 169(5):397–402
- Ganguly SS (2006) Cumulative logit models for matched pairs case-control design: studies with covariates. *J Appl Stat* 33(5):513–522
- González-Enrique J, Turias IJ, Ruiz-Aguilar JJ, Moscoso-López JA, Franco L (2019b) Spatial and meteorological relevance in NO2 estimations: a case study in the Bay of Algeciras (Spain). *Stoch Env Res Risk Assess* 33(3):801–815
- González-Enrique J, Turias IJ, Ruiz-Aguilar JJ, Moscoso-López JA, Jerez- Aragonés J, Franco L (2019c) Estimation of no2 concentration values in a monitoring sensor network using a fusion approach. *Fresenius Environ Bull* 28:681–686
- González-Enrique J, Ruiz-Aguilar JJ, Moscoso-Lopez JA, Van Roode S, Urda D, Turias IJ (2019a) A genetic algorithm and neural network stacking ensemble approach to improve no2 level estimations. In: *lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics*, 565 11506 LNCS: 856–867
- González-Enrique J, Ruiz-Aguilar JJ, Moscoso-López JA, Urda D, Turias IJ (2021) A comparison of ranking filter methods applied to the estimation of NO2 concentrations in the Bay of Algeciras (Spain). *Stoch Env Res Risk Assess* 35(10):1999–2019. <https://doi.org/10.1007/s00477-021-01992-4>
- Grivas G, Cheristanidis S, Chaloulakou A, Koutrakis P, Mihalopoulos N (2018) Elemental composition and source apportionment of fine and coarse particles at traffic and urban background locations in Athens. *Greece Aerosol Air Quality Res* 18(7):1642–1659
- Kolehmainen M, Martikainen H, Ruuskanen J (2001) Neural networks and periodic components used in air quality forecasting. *Atmos Environ* 35(5):815–825
- Moscoso-López J, Ruiz-Aguilar J, González-Enrique J, Urda D, Mesa H, Turias I (2019) Ro-ro freight prediction using a hybrid approach based on empirical mode decomposition, permutation entropy and artificial neural networks. In: *Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11734 LNAI:563
- Munoz E, Martin ML, Turias IJ, Jimenez-Come MJ, Trujillo FJ (2014) Prediction of PM10 and SO2 exceedances to control air pollution in the Bay of Algeciras, Spain. *Stoch Env Res Risk Assess* 28(6):1409–1420
- Rivera CS-Y-J-S (2015) Spatial distribution and transport patterns of NO2 in the Tijuana-San Diego area. *Atmos Pollut Res* 6(2):230–238
- Romero YV (2020) Development of a multiple regression model to calibrate a low-cost sensor considering reference measurements and meteorological parameters. *Environ Monit Assess* 192(8):1–11
- Ruiz-Aguilar JJ-E (2021) A permutation entropy-based EMD-ANN forecasting ensemble approach for wind speed prediction. *Neural Comput Appl* 33:2369–2391
- Schechtman E (2002) Odds ratio, relative risk, absolute risk reduction, and the number needed to treat– Which of these should we use? *Value Health* 5(5):431–436
- Schmidt CO, Kohlmann T (2008) When to use the odds ratio or the relative risk? *Int J Public Health* 53(3):165–167
- Sumargo B (2018) Comparing better environmental knowledge based on education and income using the odds ratio. *Procedia Comput Sci* 135:105–112
- Szabó Z (2014) *Information Theoretical Estimators Toolbox*. *J Mach Learning Res* 15:217–221
- Tripepi G, Jager KJ, Dekker FW, Wanner C, Zoccali C (2007) Measures of effect: relative risks, odds ratios, risk difference, and 'number needed to treat'. *Kidney Int* 72(7):789–791
- Turias IJ, González FJ, Martin ML, Galindo PL (2008) Prediction models of CO, SPM and SO2 concentrations in the Campo de



- Gibraltar Region, Spain: a multiple comparison strategy. *Environ Monit Assess* 143(1–3):131–146
- Vicente AB, Sanfeliu T, Jordan MM (2012) Assessment of PM10 pollution episodes in a ceramic cluster (NE Spain): Proposal of a new quality index for PM10, As, Cd, Ni and Pb. *J Environ Manag* 108:92–101
- Wang P, Chen K, Zhu S, Wang P, Zhang H (2020) Severe air pollution events not avoided by reduced anthropogenic activities during COVID-19 outbreak. *Resour Conserv Recycl* 158:104814
- Wasserman L (2004a) *All of statistics: a concise course in statistical inference* brief contents. simulation. Springer, New York
- Wasserman L (2004b) *Linear and logistic regression*. linear and logistic regression. Springer, New York
- Westmoreland EJ, Carslaw N, Carslaw DC, Gillah A, Bates E (2007) Analysis of air quality within a street canyon using statistical and dispersion modelling techniques. *Atmos Environ* 41(39):9195–9205

