Environmental Processes (2020) 7:1–21
https://doi.org/10.1007/s40710-019-00407-5

REVIEW ARTICLE

Data Mining Paradigm in the Study of Air Quality

Natacha Soledad Represa[1,2] · Alfonso Fernández-Sarría[2] · Andrés Porta[1] ·
Jesús Palomar-Vázquez[2]

Received: 31 May 2019 / Accepted: 30 October 2019 / Published online: 27 November 2019

Abstract

Air pollution is a serious global problem that threatens human life and health, as well as the environment. The most important aspect of a successful air quality management strategy is the measurement analysis, air quality forecasting, and reporting system. A complete insight, an accurate prediction, and a rapid response may provide valuable information for society's decision-making. The data mining paradigm can assist in the study of air quality by providing a structured work methodology that simplifies data analysis. This study presents a systematic review of the literature from 2014 to 2018 on the use of data mining in the analysis of air pollutant measurements. For this review, a data mining approach to air quality analysis was proposed that was consistent with the 748 articles consulted. The most frequent sources of data have been the measurements of monitoring networks, and other technologies such as remote sensing, low-cost sensors, and social networks which are gaining importance in recent years. Among the topics studied in the literature were the redundancy of the information collected in the monitoring networks, the forecasting of pollutant levels or days of excessive regulation, and the identification of meteorological or land use parameters that have the most substantial impact on air quality. As methods to visualise and present the results, we recovered graphic design, air quality index development, heat mapping, and geographic information systems. We hope that this study will provide anchoring of theoretical-practical development in the field and that it will provide inputs for air quality planning and management.

Keywords Air quality · Environmental management · Air pollution · Data mining

Highlights

- 748 papers were published between 2014 and 2018 using data mining algorithms in the analysis of air pollutants.
- 48 were selected and rigorously analyzed, from which a data mining approach was proposed for air quality management.

✉ Natacha Soledad Represa
solrepresa@quimica.unlp.edu.ar

Extended author information available on the last page of the article

- The main sources of data acquisition, the most common pre-processing techniques, the principal study methods, and different ways of viewing and presenting the results are detailed.

1 Introduction

Reducing levels of pollutants provides environmental, social and economic benefits. Annually more than 3 million deaths are attributable to ambient air pollution, which becomes the world's most considerable single environmental health risk (World Health Organization 2016). In turn, managing air quality is a crucial issue to prevent climate change from deepening (Sullivan et al. 2018). Therefore, development of public policies for monitoring and management the air pollution is urgent (Wang et al. 2017a, b; Sammarco et al. 2017).

Emissions, chemical transport and the half-life of pollutants in substances have a strong impact on chemical compounds concentrations in the air, and therefore, on the air we breathe and its quality (Wang et al. 2016). At the same time, these substances undergo chemical transformations by reactions between them or by the action of sunlight, generating new compounds. Air quality is therefore a complex phenomenon that depends on multiple natural and anthropic factors (Mayer 1999).

The most important aspect of a successful air quality management strategy is measurement analysis, air quality forecasting, and reporting system (Zhang et al. 2012). Today, sensors collect useful measurements at regular intervals generating air quality data efficiently. Their rapid analysis and reporting would provide valuable information for society's decision making (Chen et al. 2017; Sammarco et al. 2017; Yang et al. 2018b).

Data mining is a computational methodology for obtaining useful information from large data in order to discern patterns of behavior to be used in analysis and prediction (Bellinger et al. 2017; Chen et al. 2017). Data mining techniques are applied to different fields of research such as marketing, medicine, biology, engineering and social sciences (Han et al. 2011). In data mining, a data warehouse stores the data in a central repository and the search is automated, which provides faster data processing and reports visualisation production (Han et al. 2011).

In recent years, an essential advance in the way of collecting, managing and analysing data has come up, leading to a new way of understanding knowledge production. The study of air quality has not been left aside. As we will see below, numerous papers have appeared using data from different sources to obtain a more detailed picture of the situation. The publication of a large number of articles studing air quality from different perspectives requires the generation of a summary paper that synthesizes the different available tools.

This study presents a systematic review of the recent literature from 2014 to 2018 on the use of data mining in the analysis of air pollutant measurements from monitoring stations and other complementary data sources. The objective of this research is to generate a summary of the novel work done in data mining and air quality, and thus facilitate the selection of new tools for future works. The integration of large and diverse types of data offers an opportunity for a better assessment and prediction of air pollutants. This work provides elements for a complete working methodology development, from data acquisition and pre-treatment to the production of useful and understandable air quality reports.

The rest of this document is organised in sections. Section 2 presents the methodology used in the bibliographic review. Section 3 summarises the data mining approach. Also, the different steps for the treatment of air quality data from its acquisition to the generation of reports are presented in Section 3. Section 4 concludes this document with some observations on the future of the topic.

2 Research Methodology

The focus of the literature search was to identify the contributions of data mining in the analysis of air quality. To start, a cursory literature review of journal articles covering topics related to "air quality" and "data mining" were conducted. In a second phase, the works in which data mining was specifically employed in the analysis of outdoor air quality monitoring data were examined, with a particular interest on descriptive and predictive studies about atmospheric pollutant levels. As a general criterion, research whose purpose was limited to identifying sources of pollution or which only employed deterministic models were omitted.

The literature search focused solely on peer-reviewed journal articles to ensure the quality of their contributions. Articles presented at conferences were excluded because they did not bring significant technical advances on the subject. We used the descriptors ["data mining" AND "air quality"] OR ["data mining" AND "air pollutant"] OR ["data mining" AND "air pollution"] in the following databases: Elsevier (SCOPUS), IEEE Xplore, Science Direct and Taylor & Francis. The period between 2014 and 2018 was taken to assure the novelty of the works.

The research was conducted between 27 and 28 November 2018 and resulted in 748 articles. Abstracts of each article were read to assess their relevance to our research objectives and to identify duplicate articles (Kitchenham 2004; Wamba et al. 2015). At the end of this process, 48 articles were selected as being pertinent to the research objectives. Also, this paper includes other studies that present essential contributions to the subject, specificaly books or reviews that focus on data mining tools, the use of big data or data visualization. The literature review was carried out by identifying the relevant methodological contributions proposed in them. To this purpose, it was necessary to develop a framework to structure the methodologies identified by stages.

3 A Data Mining Framework for the Study of Air Quality

The short-term (public health) and long-term (climate change) benefits of controlling air quality have been studied extensively in the scientific literature (Zhang et al. 2012; Mabahwi et al. 2014). However, air quality monitoring is irregular around the globe. Numerous publications state that routine air quality monitoring is limited and that many countries lack air quality standards (Baldasano et al. 2003; Sulemana 2012; Petkova et al. 2013; Sammarco et al. 2017). One factor that could explain the low investment in air quality monitoring is that measurements of pollutants only have value concerning the knowledge they contribute to the creation and implementation of public policies that generate improvements for society (Zhang et al. 2012). Limited specific technical knowledge, inability to analyse measurements or problems in the elaboration and communication of reports are some of the difficulties encountered by public administrations (Sammarco et al. 2017; Amegah and Agyei-Mensah 2017).

The data mining paradigm can help in the study of air quality by providing a structured working methodology simplifying data analysis (Gong and Ordieres-Meré 2016). As we will see below, the process starts with the data acquisition and the data pre-processing, before applying mathematical algorithms. The methods used in data mining come from different disciplines including Artificial Intelligence, Statistics, Mathematics, Automatic Learning, and Database Systems (Han et al. 2011). Finally, a post-processing stage is used to visualise the results of the analysis in an intuitive and easy-to-communicate manner (Bellinger et al. 2017). Figure 1 proposes an architecture of the data mining approach for air quality analysis. Data sources contribute to generate databases on different variables of interest. These data are analyzed to generate useful information for end users.

Data mining algorithms are divided into description and prediction techniques. Likewise, these techniques could be separated into supervised (mostly for prediction) and unsupervised (mostly for description). Descriptive models allow to determine patterns in data sample and sub-divide them into clustering and association rules. In the predictive model, we can predict values from a different set of sample data with classification and regression algorithms (Shi et al. 2017).

To facilitate the analysis we propose a series of questions that are important in the study of air quality:

Monitoring issues

Is there redundancy of information between concentrations measured by monitoring stations? How does weather affect measured concentrations?

What are the most frequent levels of contaminants?

Does the concentration of contaminants exceed legal limits?

Analysis of extreme events

What days/months do extreme events occur?

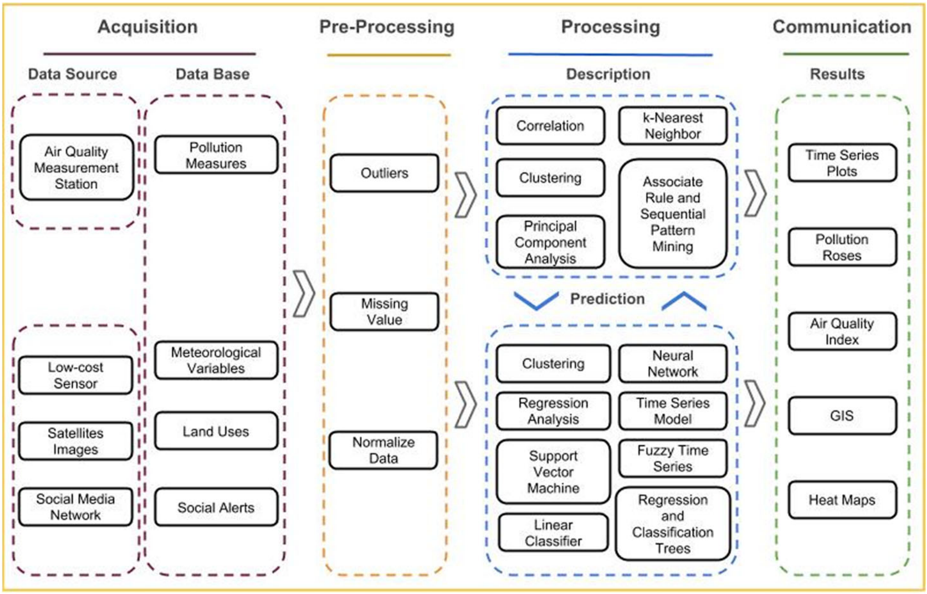

Fig. 1 The architecture of data mining approach for air quality analysis

What kind of conditions are associated with high levels of contaminants?

Temporal and spatial behavior At what time of day is the concentration of pollutants highest?
In what month do episodes of peak concentrations occur?
How have levels of pollutants changed over the years?
Where do peak concentrations occur?
What will air quality look like in the future?

The subsections below detail the different considerations and techniques identified in the bibliography for each step attempting to answer the questions presented. Whereas the first subsection oversees the analysis of different data sources and acquisition methods, the second subsection deals with data pre-processing. The third subsection describes data analysis methods, both according to descriptive and predictive models. Finally, the last subsection details how to store the results generating the best graphs and reports.

3.1 Data Sources and Acquisition

Ambient air is a complex mixture of compounds in gaseous, solid and liquid states that continuously change and interact with each other. Air quality researches study pollutants that have definite allowable concentrations in national air quality standards. Therefore, in general, air quality studies in urban areas analyse the concentration levels of the most monitored pollutants: PM_{10}, $PM_{2.5}$, NO_2, NO, O_3, CO, and SO_2. Recent works incorporate pollen as urban pollutants due to its proven effects on health (Csépe et al. 2014).

Although the data mining approach does not require massive amount of data (big data), the investigations take as their primary source the state agencies that operate continuous monitoring networks and provide hourly data (Gong and Ordieres-Meré 2016). The high frequency of pollutant measurements improves the possibility of making forecasts, in addition to a better characterisation of the phenomenon.

When this level of temporal detail is not available, a monthly analysis of contaminant concentrations may be performed. Air quality time series show seasonal behaviors produced by natural and anthropogenic factors that are interesting to consider in studies of short periods (Elangasinghe et al. 2014a). In order to represent the variability of the phenomenon in the summary measure, sampling must be done on different days and at different hours (Elangasinghe et al. 2014b).

In turn, atmospheric pollutants have high spatial variability, decreasing concentrations away from emission sources. When there is a low density of monitoring sites, the use of spatial interpolation techniques is discouraged (Qin et al. 2015; Tian et al. 2019), requiring more complex models. Air Quality Models (AQM) can give a deterministic description of air quality, including an analysis of factors and causes. AQM are based on theoretical equations that model the transport, dispersion and reactions that occur in the air after the emission of a pollutant. As a disadvantage, AQM requires specific data that is difficult to obtain everywhere (Gulia et al. 2015). We found studies that incorporate outputs from global AQM models as a source of data for statistical models, such as the LUR models detailed below (Yang et al. 2017).

It is also interesting to explore the data generated by emerging technologies (Leung et al. 2018). Remote sensing is a top-down method for obtaining daily data with global coverage on different atmospheric pollutants (Westerlund et al. 2014). On this, the scientific community has supported its use for global and mesoscale monitoring of gases and aerosols (Yeganeh et al. 2017; Bellinger et al. 2017). The literature presented an extensive use of free-download satellite images, that allow the generation of models with good accuracy and moderate spatial and temporal resolution (Rathore et al. 2015; Yang et al. 2017; Chen et al. 2018a, b).

On the other hand, there is a new group of monitoring equipment, popularly known as low-cost sensors, which for its reasonable price may be used to make dense monitoring networks (Hu et al. 2016; Honarvar and Sami 2018). This equipment is proposed to infer the air quality of a road or region that needs an estimate in real time. An interesting social aspect that appears next to this is the crowdsourcing approach, where a large group of people is asked to contribute to air quality sounding (Amegah and Agyei-Mensah 2017; Shi et al. 2017).

Due to the complex association between atmospheric pollution concentrations and environmental and meteorological conditions, the papers incorporate accessory variables that complement the research. Some works integrate meteorological variables (Gong and Ordieres-Meré 2016; Franceschi et al. 2018), environmental variables (Hasenfratz et al. 2015) or land use variables (Sadat et al. 2015; Hasenfratz et al. 2015; Honarvar and Sami 2018).

As a novel proposal, recent works integrate information from online social networks, such as messages in microblogs (Sammarco et al. 2017; Ni et al. 2017). The parameter being studied is the volume of generated messages that have certain keywords. In this way it is possible to evaluate the effectiveness of social media as a complement to air quality sensors to detect extreme events. However, public perception, awareness and response to pollution are difficult variables to quantify (Wang et al. 2015).

With the incorporation of multiple sources it is necessary to create and manage efficient and consistent databases and data warehouses (Villar et al. 2018). Data Extraction, Transformation, and Loading (ETL) systems and tools facilitate integration of data from heterogeneous sources, allowing experienced users to perform normalization, conversion, validation, and filtering operations (Castellanos et al. 2014).

3.2 Data Preparation

Data preparation is an essential step of the whole process that consumes most of the time (Bellinger et al. 2017). In this step, all inconsistencies should be corrected, such as treatment of false zero values or negative values for concentration measurements. Also, in this stage decisions must be made about the treatment of outliers and missing values replacement, which require special care because they are linked to the selection of the method of analysis to be performed.

Data can be missing due to measurement error, human error, hardware problems, insufficient sampling frequency, and faulty equipment (Junger and De Leon 2015). There are different methodologies to deal with the presence of missing data. Some works prefer to omit the stations, days or months in which a large proportion of missing data is presented. Zhang et al. (2018) calculated the 24-h mean concentration only when more than 20 h of valid data were available for the day in question. A widely used rule is to set a maximum limit of missing or erroneous values during the selected period, e.g., 15% during the selected period based on 90% of the data captured annually (European

Commission 2008). Franceschi et al. (2018) allowed 15% of missing values over the total of the measurements made. However, this limit can be extended according to a pragmatic criterion.

Data discontinuity poses an obstacle in time-series analysis, which generally requires continuous data as a condition for their use. When the proportion of missing data is not large, it is possible to fill in the missing data. One of the most commonly used methods for filling in small amounts of missing values is the linear interpolation method (Terry et al. 1986). Other more complex methods are the cubic spline interpolation, nearest neighbor interpolation or regression-based interpolation (Junninen et al. 2004; Sadat et al. 2015; Yang et al. 2018a, b). Junninen et al. (2004) showed an improvement in the performances using the multiple imputations, where the final estimate is composed of the outputs of several multivariate filling methods.

Filling techniques cannot be applied to a large amount of missing data. Junger and De Leon (2015) observed that different methods generated satisfactory results when the amount of missing data was close to 5%. However, the prediction was not sustained when the proportion of missing values exceeded 10% (Junger and De Leon 2015). Even with small amount of data (<5%), imputation using the mean or median should be avoided (Junninen et al. 2004).

Another problematic aspect is the detection of outliers. These can be easily identified when the data have a normal probability distribution. However, air quality data presents a continuous positive probability distribution, where a higher proportion of low concentrations is observed, decreasing the probability of higher concentrations. Concentrations of air pollutants usually exhibit a gamma (Bakhtiarifar et al. 2017) or lognormal (Hasenfratz et al. 2015) distribution. The asymmetric dispersion towards higher concentrations generates great difficulties when highlighting atypical events over measurement errors.

The application of limits related to the standard deviation associated with the data (e.g., control charts) cannot be used because data normality is a requirement (Martínez et al. 2014). Instead, this problem requires another perspective, which is generally based on non-parametric methods. Holesovsky et al. (2018) proposed a two-step procedure which analyses kernel smoothing residuals using extreme event theory. Martinez et al. (2014) proposed to use a functional viewpoint based on the concept of functional depth. While the method proposed by Holesovsky et al. (2018) was not able to distinguish the outliers caused by measurement and experimental errors from the outliers that result from unusual measurement conditions or from natural variability of the observed variables, Martinez et al. (2014) achieved this objective, proving this to be a more powerful approach to efficient differentiation.

When it is a requirement, a data transformation technique to stabilise the variance in order to normalise the distribution can be applied. The most widely used polynomial transformation is the Box-Cox transformation (Martinez et al. 2014).

The data preparation step includes additional data processes, such as daily or monthly aggregation of data (Desarkar and Das 2018; Gong and Ordieres-Meré 2016). The annual averages are not recommended because they hide the inter-annual changes that have great importance in the condition of the air (Zhang et al. 2018).

The mean is usually the most used aggregation function because the median presents better descriptive power when the data do not follow normal behaviour; therefore, some countries prefer to use the percentile method to assess air quality (Salako and Hopke 2012).

3.3 Data Processing

Air quality time series are nonlinear and non-stationary with different frequency characteristics challenging to be analysed (Austin et al. 2013; Chen et al. 2017). Selecting a correct analysis model is a sensitive point in data mining because it can lead to incorrect results.

The analysis methodologies identified in the literature are detailed below, taking for their presentation a distinction between descriptive and predictive methods. These methods are very diverse among themselves, some methods are supervised, and others are unsupervised (Chen et al. 2017). This arrangement is intended to provide a better understanding of the techniques within the framework of its implementation. For further details in data mining methods we recommend the studies of Gong and Ordieres-Meré (2016) and Witten et al. (2016).

3.3.1 Descriptive Methods

Several summary statistics are used to describe the set of observations. It is usual to present different measures of central tendency, such as the arithmetic mean and the median, and a measure of statistical dispersion like the standard deviation or the range (Marc et al. 2016; Chen et al. 2017).

One way to qualitatively analyse the strength of association between contaminant concentrations and related meteorological factors is to perform a correlation coefficient analysis to show which factors are most related (Yang et al. 2011; Chen et al. 2014). The most widely used correlation coefficient in the literature is the Pearson correlation coefficient (Zhang et al. 2018); in some papers, the Spearman correlation coefficient is also presented along with it (Chen et al. 2017; Ni et al. 2017).

However, when the time series are non-stationary, some statistical methods may not be entirely suitable for characterising their behavior. The multifractal approach of detrended cross-correlation analysis is a proper tool to obtain a detailed description of the relationships between two time series. The fractal approach can divide the whole data into smaller self-similar fragments and discover the physical process (Podobnik and Stanley 2008; Zhang et al. 2015; Qiao et al. 2017).

When working with many accessory variables, it is essential to establish which variables have a substantial impact on data behavior (Franceschi et al. 2018). To this end, statistical methods are usually used to choose the most appropriate variables to be incorporated into the analysis (Domańska and Łukasik 2016). Essential methods in this scope include principal component analysis (PCA), association rules mining and cluster analysis. Domańska and Łukasik (2016) present an interesting discussion regarding the selection of the dimensional reduction method. Seventeen reduction algorithms are evaluated, concluding that Isomap, Landmark Isomap, and Factor Analysis are superior since the other methods tended to remove important attribute data when eliminating redundant information.

Dimensional reduction methods may be pre-processing procedures for predictive models in order to reduce the number of input variables in the system, thus considerably diminishing redundant information, instabilities and overfitting (Russo et al. 2015). The selection of the most explanatory variables can be done independently for each station or considering all stations as a whole. Some of the techniques presented below can be employed to forecasts.

Clustering Cluster analysis algorithms cover the goal of finding different groups of similar objects. One advantage of grouping to discover the underlying structure is that it does not require human supervision, making it one of the most popular techniques (Chen et al. 2017).

Clustering methods have different uses in air quality study. Marc et al. (2016) used clustering with the objective to find links among concentrations of chemical compounds to identify potential sources of emissions in a monitored area. Another problem solved with clustering methods was grouping measurement points or observation objects in different sampling seasons (Marc et al. 2016; Xie et al. 2018). Cluster methods can be applied to the measured concentration matrices or on the correlation matrix according to the goal (Xie et al. 2018; Wang and Zhao 2018). At the same time, the numbers and the significance of the clusters reflect the strength of the relationship between two data sets.

This technique has been used to identify redundant data between monitoring stations. He et al. (2018) made an analysis of the Spearman correlation factor and applied a cluster analysis in order to reveal the similarities of the $PM_{2.5}$ monitoring network behaviour in Shanghai.

In connectivity-based clustering, clusters are formed by connecting data points according to their distance. These methods are commonly referred to as "hierarchical grouping" because they produce a hierarchy partition of the dataset from which the user selects the appropriate level of clusters. These methods are not very robust towards outliers. We can see an application in Wang et al. (2018) where a cluster analysis was conducted using seasonal values to identify regions with a similar variation in contaminant concentration.

In the Centroid-based clustering model, clusters are represented by a centroid, which is not necessarily a member of the data set. The distance between a data point to the centroid indicates the similarity between the two objects. This algorithm requires careful handling of time series of unequal lengths and is highly susceptible to noise and outliers (Yang et al. 2018a).

An important step is to establish which distance measure is more appropriate for the characteristics of the time series analysed. The most commonly used methods of this model are k-means and k-Medoids. Elangasinghe et al. (2014a, b) used a k-means data mining technique in polar diagrams to find similarities and relationships among wind direction, wind speed, and PM_{10} and $PM_{2.5}$ concentrations to identify the primary sources of emissions in New Zealand. Chen et al. (2015) suggested an Environmental Pollution Clustering algorithm to organise a large number of sample data points into groups based on k-means, with the advantage that it does not impose a fixed number of clusters.

In general, it is possible to obtain better results using an approach that contemplates temporal correlation of the data. In turn, air pollutants time series may be a noisy environment. Consequently, ignoring time dependence of the data could produce results biased by the presence of outliers (Mori et al. 2016; Chen et al. 2017). Some methods are based on replacing the distance/similarity measure for static data with an appropriate measure for time series. Another approach is to reduce time series to a smaller vector or model parameters, and then applying a conventional grouping algorithm to the vectors or model parameters (Liao 2005). It has been observed that the performances of robust fuzzy models are better than results obtained utilising standard (non-fuzzy) and non-robust clustering procedures based on hierarchical and partitioning around centroid (e.g., k-means clustering) approaches (Chen et al. 2017).

K-Nearest Neighbour The k-nearest neighbors' algorithm (k-NN) is a non-parametric method employed for the classification and regression of a dataset. Objects are categorised according to the most common class among their k nearest neighbors. In the regression of k-NN, in contrast, the output is the average of the values of its closest neighbors to k. This technique can be used to assign weight to neighbors' contributions based on distance, so that the nearest neighbors contribute more to the average than the furthest (Junninen et al. 2004). In this sense, k-NN techniques include the spatial-temporal relationship in order to classify the data. As an example of its integration with other techniques, Soh et al. (2018) developed an Adaptive Deep Learning-based Air Quality Prediction Model using k-NN by Euclidean Distance and k-NN by Dynamic Time Warping (DTW) Distance.

Associative Rule and Sequential Pattern Mining Association Rule Mining (ARM) seeks association among a large set of categorical data. An association rule shows and enables to predict the occurrence of a specific event based on the occurrences of the other events (Sadat et al. 2015; Witten et al. 2016). A variant of ARM is the Sequential Pattern Mining (SPM). SPM was developed to discover the associative relationships between lists of data records ordered according to either spatial or temporal dimensions (Yang et al. 2018b). Although ARM and SPM are usually represented by a rule with an antecedent and a consequence, it should be differentiated from a causal analysis. While the ARM and SPM indicate a statistical relationship between the observations, a causal relationship implies whether the presence of one incident causes the presence of another (Soysal 2015; Yang et al. 2018b).

Fuzzy logic provides a good alternative for assessments based on a reasoning process. Fuzzy ARM provides more reliable associations rules because fuzzy sets can model the uncertainty embedded in the measures (Olvera-García et al. 2016). The fuzzy analysis takes particular importance for spatial association rule mining where the uncertainty is higher (Sadat et al. 2015). An application case is (Jiang et al. 2017), where an association rule was introduced to determine the associated cities that have strong relativity around the target domain.

Principal Component Analysis Principal Component Analysis (PCA) is a feature extraction method that uses orthogonal linear projections to capture the underlying variance of the data. The dataset with interrelated variables is transformed into a reduced new set of variables which are the combinations with the highest variance (Franceschi et al. 2018).

PCA has been successfully applied for AQ studies as a non-parametric method of classification. PCA has been used to relate meteorological variables to the concentration of atmospheric pollutants (Franceschi et al. 2018) and to classify the monitoring sites (Pires et al. 2008). This technique is useful for optimising the monitoring sites that constitute the network, helping to reduce costs, but at the same time ensuring adequate characterisation of regional air quality. A novel application of this technique can be found in Harkat et al. (2018), where PCA and interval-valued PCA were used to predict unmeasured variables from others; it was easier to measure and useful for fault detection in air quality data.

Linear Classifier A linear classifier is a classifier which uses a hyperplane to separate classes. A linear classifier is robust against noise since it tends not to overfit and has relatively shorter training and testing times (Shmilovici 2009). Two large broad classes of methods are used to determine the parameters of a linear classifier. The generative methods have as an example: Linear Discriminant Analysis that assumes Gaussian conditional density models, and Naive Bayes Classifier with multinomial or multivariate Bernoulli event models, while discriminative models which cover the Logistic Regression, Perceptron, and SVM are also present.

SVM models are a set of related methods for supervised learning, applicable to classification, regression, and non-linear function approximation. Its employment is recommended for having a better generation capability, global optimisation, and dimensional independence (Shmilovici 2009; Xu et al. 2017). Also, empirical comparisons have shown that SVM performs faster in the training set than Artificial Neural Networks (ANN), even though ANN is the most robust to work with noise (Gong and Ordieres-Meré 2016). Xu et al. (2017) compared a hybrid forecasting model (SVM with a Whale Optimization Algorithm) to ARMA model; the proposed model proved to be superior in three cities.

3.3.2 Predictive Models

Forecasting pollution levels are an essential aspect of air quality management. The various approaches proposed for predicting atmospheric concentrations are primarily divided into deterministic and stochastic models.

Deterministic or numerical models, including dispersion and chemical transport models, have various problems and deficiencies. As mentioned above, appropriate inputs are difficult to determine or they have a high degree of relative uncertainty because they are sensitive to a large number of factors, including the scale and quality of the source inventories, the processes of dispersion and distribution, and chemical reactivity into the atmosphere (Wang et al. 2017a, b). In turn, hardware requirements are often high for the implementation of traditional models (Jiang et al. 2017). These requirements make it difficult to adopt a single prediction model for air quality management.

Stochastic or empirical observation-based methods rely on the relationship between current air quality measurements and historical measurements, and may also include different variables. The conventional statistical approaches in prediction include linear or nonlinear regression models, time series models (ARMA, ARIMA, SARIMA, Holt-Winters) and support vector machine (SVM) (Wang et al. 2017a, b). Its approach usually requires a large amount of historical data to build the mapping between the predictors and targets, making the applicability of statistical models limited (Jiang et al. 2017). Innovative approaches for air quality forecasting include stepwise (STW) regression and wavelet analysis (Russo et al. 2015), artificial neural networks (Csépe et al. 2014; Sekar et al. 2015; Gong and Ordieres-Meré 2016), fuzzy and neuro-fuzzy logic (Gong and Ordieres-Meré 2017; Dincer and Akkuş 2018), and hidden Markov models (Domańska and Łukasik 2016; Gómez-Losada 2017; Jiang et al. 2017).

As noted above, descriptive data mining techniques can also be applied in prediction. For example, clustering analysis is also used for forecasting (Domańska and Wojtylak 2014). The overall cluster forecast can be obtained through a linear regression between the centre of the cluster and the total cluster. At first, the cluster centre is estimated with the historical data and, then the forecast made is extended to the entire cluster. This approach can help when variability is significantly affected by causal factors that can be identified and monitored (Zotteri et al. 2005).

Different works use diverse statistical measures to evaluate the performance of the models, such as the correlation coefficient (R), determination coefficient (R^2), the normalized mean square error (NMSE), the root mean square error (RMSE) or the mean absolute error (MAE) (Csépe et al. 2014; Russo et al. 2015). Some studies incorporated another measure of forecasting error, such as the fractional bias, the factor of two (Biancofiore et al. 2017), the Nash–Sutcliffe efficiency index (Sekar et al. 2015) or accuracy (Wang and Song 2018).

Regression Analysis Forecasts using the linear regression method assume that the relationship between the two variables is constant over time. The permanence of the previously established relationship is fundamental for the forecasting of the response variable. A generalisation of linear regression is the Generalized Linear Model (GLM), which allows treating error distribution models other than a normal distribution (Westerlund et al. 2014; Hasenfratz et al. 2015; Biancofiore et al. 2017).

Regression linear analysis can be expanded to include more predictive variables. Land Use Regression (LUR) models use a set of explanatory variables to model pollution concentrations (Westerlund et al. 2014; Shi et al. 2017b; Yang et al. 2017). In scientific literature, LUR models can be found based on Generalized Linear Mixed Models (GLMMs) or Generalized Additive Models (GAMs) (Hasenfratz et al. 2015).

In the Generalised Additive Model (GAM), the linear predictor depends linearly on unknown smooth functions of a predictor variable (Hastie 2017). GAMs have been used to analyse and model the spatiotemporal variability of a particulate matter because it can adjust for non-linear confounding parameters such as seasonal changes, trends, and the weather variables (Hasenfratz et al. 2015). Nevertheless, its most significant application is in the field of epidemiology, with models that relate contaminant levels to health effects (Lin et al. 2016; Wu et al. 2016). GLMMs include random effects in the linear predictor, giving an explicit probability model that explains the origin of the correlations. Yang et al. (2018a) presented a mixed effects model using MODIS 3 km AOD together with meteorological data to estimate $PM_{2.5}$ concentrations.

When working with multivariate models, it is practical to use the automatic STW regression procedure instead of manually conducting an exploratory regression model. The widely used STW regression technique is an iterative regression modeling method that automatically selects independent variables by testing the statistical significance of regression models. This procedure automatically adds and removes predictors to find a subset of variables that, together with the previously selected variables, generate the prediction more accurately (Russo et al. 2015; Shi et al. 2017b).

Least Absolute Shrinkage and Selection Operator (LASSO) is another useful method used to select the most relevant predictive variables for linear regression. LASSO is based on *minimising* Mean Squared Error by the addition of a penalty term on the parameter vector (Li and Shao 2015; Yeganeh et al. 2017). Prediction based on regression analysis requires that air pollution datasets satisfy some statistical assumptions. Since air pollution datasets present seasonality and are highly dispersed, fulfilling these assumptions is very difficult (Dincer and Akkuş 2018). For this reason, in many studies, non-parametric techniques are preferred for forecasting.

Box Jenkins Models Regression Analysis and Box-Jenkins models lead among statistical techniques used for predicting future values (Sharma et al. 2009; Dincer and Akkuş 2018). The Box-Jenkins Model is a systematic method of identifying, fitting, checking, and using integrated ARIMA time series models. Ni et al. (2017) used ARIMA model to establish a correlation analysis model of $PM_{2.5}$ to meteorological data and social media data.

Fuzzy Time Series When the data is incomplete, includes uncertainty and does not satisfy statistical assumptions as normality, a strategy is to transform data into fuzzy data. As mentioned above, fuzzy sets are used to represent interval events in the domains of continuous attributes, allowing continuous data lying on the interval boundaries to belong to multiple intervals partially. Fuzzy Time Series (FTS) are appropriate for forecasting air pollution data with the time series approach (Dincer and Akkuş 2018). The primary disadvantage is that the majority of fuzzification methods are sensitive to outliers. Dincer and Akkuş (2018) proposed a new fuzzy time series model using Fuzzy K-Medoids clustering algorithm in data sets with outliers and noise data points.

Regression and Classification Tree Decision tree learning uses a decision tree as a model by performing binary recursive partitioning. Initially, the data are split into subsets based on different evaluation functions (Gini diversity index, the entropy, or the error index) (Gacquer et al. 2011). Then, the method subjects each subset to the same partitioning process until only a few samples remain in the terminal subset (Gong and Ordieres-Meré 2016). The objective is to create a predictive model by learning simple decision rules inferred from the data features (Desarkar and Das 2018).

Decision tree techniques are implemented for continuous or discrete variables. For Regression Tree, the target variable takes continuous values, whereas, for Classification Tree, the variable takes a discrete set of values. The Classification and Regression Trees (CART) are easily comprehensible because they provide the reason for the chosen solution. Also, CART is a non-parametric data-driven technique, i.e., the number of parameters is not specified before modeling. Another advantage of CART is its simplicity and efficiency of their construction compared to other classifiers such as Neural Networks (Desarkar and Das 2018) and Support Vector Machines (Gong and Ordieres-Meré 2016).

A variety of algorithms based on CART can be found. Csépe et al. (2014) used different Tree-based algorithms (M5P, J48, C4.5, REPTree, DecisionStump) to forecasting ragweed pollen concentration. Sekar et al. (2015) used the M5P Model and REPTree to predict $PM_{2.5}$ and CO concentrations (Witten et al. 2016).

Random Forest Random Forest (RF) algorithm is CART based. RF is an ensemble algorithm that applies to bagging methods on CART. Unlike CART, RF selects only a few features randomly in each training time. In this way, RF can obtain a lower error level and running time without using all the input variables (Birant 2011). The performance function is a measurement of the models' performance (Gong and Ordieres-Meré 2016). Gong and Ordieres-Meré (2016) employed CART and RF to identify the relevant variables. Zhao and Song (2017) used RF with logistic regression and linear discriminant analysis to predict $PM_{2.5}$ concentration in different air conditions.

Artificial Neural Network and Deep Learning ANNs are becoming the widely used and useful alternative techniques to model non-linear systems such as air pollution (Fu et al. 2015; Biancofiore et al. 2017; Jiang et al. 2017; Ni et al. 2017; Russo et al. 2015). Deep learning algorithms are multi-layered *neural ANNs* with which it is possible to perform recurrent analysis or feedforward for the extraction of hierarchical characteristics (Wang and Song 2018). These models represent a good compromise between flexibility and effectiveness, do not require any statistical assumptions or constraints, and may provide better forecasting performance than statistical models (Sekar et al. 2015; Russo et al. 2015; Franceschi et al. 2018). However, ANN models have some drawbacks as local minima problems, poor generalisation issues, difficulty in selecting appropriate network architecture, and complexity of computation (Dincer and Akkuş 2018).

Hybrid models combine individual models for better air quality forecasting. ANN models are typically combined with other models to obtain more accurate air pollutant concentration forecasting

(Jiang et al. 2017). Gong and Ordieres-Meré (2016) used a feed-forward back propagation ANN and the Pre-processing Synthetic minority over-sampling technique (SMOTE) to adjust the sampling balance and prevent the overfitting problem. Csépe et al. (2014) applied two types of neural networks: a complex (Multi-Layer Perception with more than one hidden layer) and a less complex (Multi-Layer Perception Regressor with only one hidden layer) version. Other researchers designed a fully connected deep LSTM network as a temporal predictor (Wang and Song 2018).

Jiang et al. (2017) used an adaptive fuzzy neural network (AFNN) to make $PM_{2.5}$ concentration predictions. Yeganeh et al. (2017) employed a wide range of ground-based $PM_{2.5}$ measurements, land use, meteorological and remotely-sensed AOD data to estimate the $PM_{2.5}$ concentration using adaptive neuro-fuzzy inference system (ANFIS), SVM and back-propagation artificial neural network (BPANN) algorithms.

While some papers take the complete time series to feed the model, in other cases a random sample is taken from the time. Randomised Data ensure the robustness of the trained network by providing normally distributed samples for training, cross-validation and testing (Elangasinghe et al. 2014a, 2014b). Elangasinghe et al. (2014a) shuffled the time series of one complete year of data, and the sample size was 300 data points for each season.

A large number of methods have been identified in this literature review. However, the authors present different error measures making their comparison problematic. As a summary, Table 1 shows the performance of various prediction models. It seems that the neural networks presented an advantage over the other methods, although there is not enough bibliography to support it.

3.4 Data Visualization

The final stage in data management is the elaboration of reports and visualisations to highlight the results obtained in the previous steps. Even if data mining is highly

Table 1 Comparison of forecasting model performance (Modified from Franceschi et al. 2018)

Model description	Air pollutant	Location	RMSE ($\mu g\ m^{-3}$)	Authors
Daily average concentration by ANN and air mass trajectory model	$PM_{2.5}$	China	15.65	Feng et al. (2015)
Daily average concentration by Back Propagation ANN based on wavelet decomposition.	PM_{10}	China	15.39	Bai et al. (2016)
Daily average concentration by ANN-MLP with RPROP training, combined with k-means results	PM_{10} PM_{25}	Colombia	13.50 5.77	Franceschi et al. (2018)
Daily average concentration by multiple linear regression	PM_{10}	Portugal	12.8	Russo et al. (2015)
Hourly averages concentration by ANN and clustering	PM_{10} $PM_{2.5}$	New Zealand	6.34 4.74	Elangasinghe et al. (2014a)
Hourly average concentration by multiple linear regression Hourly average concentration by ANN	NO_2	New Zealand	15.79 7.07	Elangasinghe et al. (2014b)
Daily average concentration by hybrid method (improved complete ensemble empirical mode decomposition and whale optimisation algorithm and SVM)	$PM_{2.5}$ PM_{10} NO_2	China	3.99 7.16 2.41	Xu et al. (2017)

automated, humans play a central role in the result interpretation. Poor presentation of conclusions may hide relevant points, and may bias the air quality public perceptions (Ma et al. 2012). It is important to remember that this is the only part of the analytical process that reaches the audience, and therefore, must answer the central questions of interest to the public.

Effective communication requires identifying specific points of the analysis that are desired to be shown. In this sense, exploratory graphics elaborated to understand the most profound relationships of the dataset need to be left aside. Also, the visualisations designed for the exclusive purpose of facilitating the understanding of the analysis performed should be made in previous steps (Knaflic 2015).

The visual representation to be selected should provide a clear picture of the patterns and relationships discovered in order to facilitate the interpretation and understanding of the information by the audience. It is advisable to enhance the comprehension of the information that the presentation of the data can present as storytelling, with appropriate visual displays (Ma et al. 2012). Some examples of the last are given below.

The most used graph to visualise relationships is the dendrogram. It is particularly useful for presenting the layout of clusters produced by techniques of hierarchical grouping. When it is desired to show the grouping of monitoring stations, it may be convenient to present it on a map for spatial representation (Xie et al. 2018; Wang and Zhao 2018).

Time series are used to view the time pattern. It is common to aggregate data to show behaviour over weeks and months. Also, it is useful to show the complete time series with the trend of the period analysed (Elangasinghe et al. 2014a). To allow daily, monthly and annual comparisons of air pollution variability it is convenient to employ a calendar chart for an easy visualisation (Leung et al. 2018).

Pollution roses represent the contamination levels observed by each monitoring station according to wind direction. The pollution rose model is based on the simple idea of classifying the concentration measurements obtained at a given site for a given period according to the direction in which the wind blows at the moment of the measurement (Duboue 1978). It provides a relative notion of pollution sources location.

Maps are an effective way to visualise and communicate the spatial variation of pollutant concentrations. Heat maps are used to identify higher concentration areas (Austin et al. 2013). However, if an appropriate color scale is not selected, they become confusing.

In order to facilitate public interpretation, indices are often developed to report the results (Zhang et al. 2012). The Air Quality Index (AQI) is a dimensionless index for reporting daily air quality to the general population. The AQI converts a pollutant concentration to a number on a scale of 0–500, and the air quality is characterised by “very high”, “high”, “moderate” and “low” (Sadat et al. 2015). The categorisation breakpoints used vary from one air pollutant to another (Mintz 2012). A variant can be observed in Olvera-Garcia et al. (2016) where they applied a diffuse analysis to generate AQI weights for contaminants with significant health effects.

Regarding the tools employed, the bibliography analysis shows a variety of computer software use. The *openair* R package is a convenient tool for the evaluation of air pollutant measurements (Carslaw and Ropkins 2012). Also, several studies included the use of geographic information systems (GIS) to generate spatial analyses and visualisations (Alsahli and Al-Harbi 2018; Shahbazi et al. 2016; Liu et al. 2017).

Finally, to strengthen the narrative of the presentation of results, it is convenient to compare the measured or predicted concentrations with air quality standards. Since, generally, there are differences between the air quality standards and the World Health Organization (WHO) recommendations, it is appropriate to compare them with both values. Economic policy issues often condition air quality standards and take a long time to update, while WHO recommendations are based on scientific researches that support these values and are also revised when sufficient evidence is gathered (Zhang et al. 2018).

4 Conclusions

Air pollution is a major global problem that threatens human life and health as well as the environment. An enormous amount of data from different sources is presented in the form of raw data and contains relevant information that can be used to improve monitoring, build AQM or generate predictions. A general systematic methodology is needed to integrate, analyse and communicate these data, along with the design of scalable architectures to contain new sources in the future.

In recent years, data mining has contributed with many techniques and technologies in this field, although its adoption is still early. This article summarises the most relevant developments in the area over the years 2014–2018. For this purpose, a data mining architecture for air quality analysis has been proposed, consistent with the literature analysed. To achieve effective air quality management, data acquisition, pre-treatment and analysis and presentation of results were given the same degree of importance.

Regarding data sources, the primary sources of air quality data are oficial continuous monitoring networks. Multivariable models include covariable meteorological data, data collected by social networks or by remote sensing and also AQM outputs. The methodologies analyzed offer the possibility of integrating information from various sources to obtain better predictions of pollutants. Furthermore, the identification of meteorological or land use parameters that most affect air quality was identified as a relevant aspect. Further development of these investigations may contribute to an improvement in support systems of air quality management.

Concerning the evaluation of existing monitoring networks, assessing the information collected and evaluating its redundancy are essential to optimise its performance. Different studies have shown that cluster analysis and PCA techniques could be used for this purpose.

In general, the studies analyzed focused on predicting future contaminant levels and, in a lesser case, completing information in unmonitored areas. These publications explored techniques to obtain the best model, using different tools and without revealing a preference. However, it has become evident that the analysis of time series requires several considerations according to the methodology chosen.

In this sense, non-parametric statistical techniques such as machine learning algorithms offer great advantages. Within this group, it is worth highlighting the best prediction that hybrid models present over simple ones.

Finally, this paper presents a summary of the different ways of visualising information related to air quality. The preferred approaches to generate storytelling that is easily understood for the general public have been the use of an air quality index and the implementation of thematic maps in GIS. Nonetheless, this is a scarcely explored field that needs to grow for a better appropriation of the problem by society.

We expect that this study will provide an anchorage of theoretical-practical development in the subject, motivating new research and management projects that will allow a better decision-making process, and therefore, generate actions and interventions aimed at minimising air pollution risks, improving air quality and people's well-being. This work is also expected to contribute to the generation of health warnings and alerts for sensitive groups and the general public to the planning of urban health care and the improvement of public policies on disease prevention.

Acknowledgments The authors gratefully acknowledge the support of the Argentinean Scientific and Technical Research Council, the Polytechnic University of Valencia, Spain, and the National University of La Plata, Argentina. National Agency for Scientific and Technological Promotion funded this project through the PICT 2015-0618.

References

Alsahli MM, Al-Harbi M (2018) Allocating optimum sites for air quality monitoring stations using GIS suitability analysis. Urban Clim 24:875–886

Amegah AK, Agyei-Mensah S (2017) Urban air pollution in sub-Saharan Africa: time for action. Environ Pollut 220:738–743

Austin E, Coull BA, Zanobetti A, Koutrakis P (2013) A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition. Environ Int 59:244–254

Bai Y, Li Y, Wang X, Xie J, Li C (2016) Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. Atmos Pollut Res 7(3):557–566

Bakhtiarifar MH, Bashiri M, Amiri A (2017) Optimization of problems with multivariate multiple functional responses: a case study in air quality. Commun Statist Simul Comput 46(10):8049–8063

Baldasano JM, Valera E, Jimenez P (2003) Air quality data from large cities. Sci Total Environ 307:141–165

Bellinger C, Jabbar MSM, Zaïane O, Osornio-Vargas A (2017) A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health 17(1):907

Biancofiore F, Busilacchio M, Verdecchia M, Tomassetti B, Aruffo E, Bianco S et al (2017) Recursive neural network model for analysis and forecast of PM10 and PM2.5. Atmos Pollut Res 8(4):652–659

Birant D (2011) Comparison of decision tree algorithms for predicting potential air pollutant emissions with data mining models. J Environ Inform 17(1)

Carslaw DC, Ropkins K (2012) Openair—an R package for air quality data analysis. Environ Model Softw 27: 52–61

Castellanos MG, Dayal U, Simitsis A, Wilkinson WK (2014). Quality-driven ETL design optimization 2014. U.S. Patent No. 8:719–769. U.S. Patent and Trademark Office, Washington, DC

Chen G, Li S, Knibbs LD, Hamm NAS, Cao W, Li T, Guo J, Ren H, Abramson MJ, Guo Y (2018a) A machine learning method to estimate PM 2.5 concentrations across China with remote sensing, meteorological and land use information. Science of the Total Environment 636:52-60

Chen G, Wang Y, Li S, Cao W, Ren H, Knibbs LD, Abramson MJ, Guo Y (2018b) Spatiotemporal patterns of PM10 concentrations over China during 2005–2016: A satellite-based estimation using the random forests approach. Environmental Pollution 242:605-613

Chen J, Xin J, An J, Wang Y, Liu Z, Chao N, Meng Z (2014) Observation of aerosol optical properties and particulate pollution at background station in the Pearl River Delta region. Atmos Res 143:216–227

Chen M, Wang P, Chen Q, Wu J, Chen X (2015) A clustering algorithm for sample data based on environmental pollution characteristics. Atmos Environ 107:194–203

Chen Y, Wang L, Li F, Du B, Choo KKR, Hassan H, Qin W (2017) Air quality data clustering using EPLS method. Inform Fusion 36:225–232

Csépe Z, Makra L, Voukantsis D, Matyasovszky I, Tusnády G, Karatzas K, Thibaudon M (2014) Predicting daily ragweed pollen concentrations using computational intelligence techniques over two heavily polluted areas in Europe. Sci Total Environ 476:542–552

Desarkar A, Das A (2018) Implementing decision tree in air pollution reduction framework. In: Smart computing and informatics. Springer, Singapore, pp 105–113

Dincer NG, Akkuş Ö (2018) A new fuzzy time series model based on robust clustering for forecasting of air pollution. Ecol Inform 43:157–164

Domańska D, Łukasik S (2016) Handling high-dimensional data in air pollution forecasting tasks. Ecol Inform 34:70–91

Domańska D, Wojtylak M (2014) Explorative forecasting of air pollution. Atmos Environ 92:19–30

Duboue M (1978) Pollution roses: a simple way of interpreting the data obtained by air pollution measurement systems in the proximity of refineries. Stud Environ Sci:133–136

Elangasinghe MA, Singhal N, Dirks KN, Salmond JA (2014b) Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis. Atmos Pollut Res 5(4):696–708

Elangasinghe MA, Singhal N, Dirks KN, Salmond JA, Samarasinghe S (2014a) Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering. Atmos Environ 94:106–116

European Commission (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. Off J European Union

Feng X, Li Q, Zhu Y, Hou J, Jin L, Wang J (2015) Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmos Environ 107:118–128

Franceschi F, Cobo M, Figueredo M (2018) Discovering relationships and forecasting PM10 and PM2.5 concentrations in Bogotá, Colombia, using artificial neural networks, principal component analysis, and k-means clustering. Atmos Pollut Res 9(5):912–922

Fu M, Wang W, Le Z, Khorram MS (2015) Prediction of particulate matter concentrations by developed feed-forward neural network with rolling mechanism and gray model. Neural Comput Appl 26(8):1789–1797

Gacquer D, Delcroix V, Delmotte F, Piechowiak S (2011) Comparative study of supervised classification algorithms for the detection of atmospheric pollution. Eng Appl Artif Intell 24(6):1070–1083

Gómez-Losada Á (2017) Clustering air monitoring stations according to background and ambient pollution using hidden Markov models and multidimensional scaling. In: Data science. Springer, Cham, pp 123–132

Gong B, Ordieres-Meré J (2016) Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: case study of Hong Kong. Environ Model Softw 84:290–303

Gong B, Ordieres-Meré J (2017) Reconfiguring existing pollutant monitoring stations by increasing the value of the gathered information. Environmental Modelling & Software 96:106-122

Gulia S, Nagendra SS, Khare M, Khanna I (2015) Urban air quality management-a review. Atmos Pollut Res 6(2):286–304

Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier, New York

Harkat MF, Mansouri M, Nounou M, Nounou H (2018) Enhanced data validation strategy of air quality monitoring network. Environ Res 160:183–194

Hasenfratz D, Saukh O, Walser C, Hueglin C, Fierz M, Arn T et al (2015) Deriving high-resolution urban air pollution maps using mobile sensor nodes. Pervasive Mobile Comput 16:268–285

Hastie TJ (2017) Generalized additive models. In: Statistical models in S. Routledge, Boca Raton, pp 249–307

He HD, Li M, Wang WL, Wang ZY, Xue Y (2018) Prediction of PM2. 5 concentration based on the similarity in air quality monitoring network. Building and Environment 137:11-17

Holešovský J, Čampulová M, Michálek J (2018) Semiparametric outlier detection in nonstationary times series: case study for atmospheric pollution in Brno, Czech Republic. Atmos Pollut Res 9(1):27–36

Honarvar AR, Sami A (2019) Towards sustainable smart city by particulate matter prediction using urban big data, excluding expensive air pollution infrastructures. Big Data Res 17:56–65

Hu Y, Fan J, Zhang H, Chen X, Dai G (2016) An estimated method of urban PM2. 5 Concentration distribution for a mobile sensing system. Pervasive Mobile Comput 25:88–103

Jiang P, Dong Q, Li P (2017) A novel hybrid strategy for PM2. 5 concentration analysis and prediction. J Environ Manag 196:443–457

Junger WL, De Leon AP (2015) Imputation of missing data in time series for air pollutants. Atmos Environ 102:96–104

Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M (2004) Methods for imputation of missing values in air quality data sets. Atmos Environ 38(18):2895–2907

Kitchenham B (2004) Procedures for performing systematic reviews. Keele UK Keele Univ 33(2004):1–26

Knaflic CN (2015) Storytelling with data: a data visualization guide for business professionals. Wiley

Leung Y, Leung KS, Wong MH, Mak T, Cheung KY, Lo LY et al (2018) An integrated web-based air pollution decision support system–a prototype. Int J Geogr Inform Sci:1–28

Li Q, Shao J (2015) Regularizing lasso: a consistent variable selection method. Stat Sin:975–992

Liao TW (2005) Clustering of time series data—a survey. Pattern Recogn 38(11):1857–1874

Lin H, Liu T, Xiao J, Zeng W, Li X, Guo L et al (2016) Quantifying short-term and long-term health benefits of attaining ambient fine particulate pollution standards in Guangzhou, China. Atmos Environ 137:38–44

Liu Z, Xie M, Tian K, Gao P (2017) GIS-based analysis of population exposure to PM2. 5 air pollution—a case study of Beijing. J Environ Sci 59:48–53

Ma KL, Liao I, Frazier J, Hauser H, Kostis HN (2012) Scientific storytelling using visualization. IEEE Comput Graph Appl 32(1):12–19

Mabahwi NAB, Leh OLH, Omar D (2014) Human health and wellbeing: human health effect of air pollution. Procedia Soc Behav Sci 153:221–229

Marć M, Bielawska M, Simeonov V, Namieśnik J, Zabiegała B (2016) The effect of anthropogenic activity on BTEX, NO2, SO2, and CO concentrations in urban air of the spa city of Sopot and medium-industrialized city of Tczew located in North Poland. Environ Res 147:513–524

Martínez J, Saavedra Á, García-Nieto PJ, Piñeiro JI, Iglesias C, Taboada J et al (2014) Air quality parameters outliers detection using functional data analysis in the Langreo urban area (Northern Spain). Appl Math Comput 241:1–10

Mayer H (1999) Air pollution in cities. Atmos Environ 33(24–25):4029–4037

Mintz D (2012). Technical assistance document for the reporting of daily air quality-the air quality index (aqi): US environmental protection agency. Office of Air Quality Planning and Standards

Mori U, Mendiburu A, Lozano JA (2016) Similarity measure selection for clustering time series databases. IEEE Trans Knowl Data Eng 28(1):181–195

Ni XY, Huang H, Du WP (2017) Relevance analysis and short-term prediction of PM2.5 concentrations in Beijing based on multi-source data. Atmos Environ 150:146–161

Olvera-García MÁ, Carbajal-Hernández JJ, Sánchez-Fernández LP, Hernández-Bautista I (2016) Air quality assessment using a weighted fuzzy inference system. Ecol inform 33:57–74

Petkova EP, Jack DW, Volavka-Close NH, Kinney PL (2013) Particulate matter pollution in African cities. Air Qual Atmos Health 6(3):603–614

Pires JCM, Sousa SIV, Pereira MC, Alvim-Ferraz MCM, Martins FG (2008) Management of air quality monitoring using principal component and cluster analysis—Part I: SO2 and PM10. Atmos Environ 42(6):1249–1260

Podobnik B, Stanley HE (2008) Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series. Phys Rev Lett 100(8):084102

Qiao ZX, Pan W, Lu WZ (2017) Multiscale multifractal properties between ground-level ozone and its precursors in rural area in Hong Kong. J Environ Manag 196:270–277

Qin S, Liu F, Wang C, Song Y, Qu J (2015) Spatial-temporal analysis and projection of extreme particulate matter (PM10 and PM2.5) levels using association rules: A case study of the Jing-Jin-Ji region, China. Atmospheric Environment 120:339-350

Rathore MMU, Paul A, Ahmad A, Chen BW, Huang B, Ji W (2015) Real-time big data analytical architecture for remote sensing application. IEEE J Sel Top Appl Earth Obs Remote Sens 8(10):4610–4621

Russo A, Lind PG, Raischel F, Trigo R, Mendes M (2015) Neural network forecast of daily pollution concentration using optimal meteorological data at synoptic and local scales. Atmos Pollut Res 6(3):540–549

Sadat YK, Nikaein T, Karimipour F (2015) Fuzzy spatial association rule mining to analyze the effect of environmental variables on the risk of allergic asthma prevalence. Geodesy Cartogr 41(2):101–112

Salako GO, Hopke PK (2012) Impact of percentile computation method on PM 24-h air quality standard. J Environ Manag 107:110–113

Sammarco M, Tse R, Pau G, Marfia G (2017) Using geosocial search for urban air pollution monitoring. Pervasive Mobile Comput 35:15–31

Sekar C, Gurjar BR, Ojha CSP, Goyal MK (2015) Potential assessment of neural network and decision tree algorithms for forecasting ambient PM 2.5 and CO concentrations: case study. J Hazard Toxic Radioactive Waste 20(4):A5015001

Shahbazi H, Taghvaee S, Hosseini V, Afshin H (2016) A GIS based emission inventory development for Tehran. Urban Clim 17:216–229

Sharma P, Chandra A, Kaushik SC (2009) Forecasts using box–Jenkins models for the ambient air quality data of Delhi City. Environ Monit Assess 157(1–4):105–112

Shi D, Guan J, Zurada J, Manikas A (2017) A data-mining approach to identification of risk factors in safety management systems. J Manag Inf Syst 34(4):1054–1081

Shi Y, Lau KKL, Ng E (2017b) Incorporating wind availability into land use regression modelling of air quality in mountainous high-density urban environment. Environ Res 157:17–29

Shmilovici A (2009) Support vector machines. In: Maimon O, Rokach L (eds) Data mining and knowledge discovery handbook. Springer, Boston, MA

Soh PW, Chang JW, Huang JW (2018) Adaptive deep learning-based air quality prediction model using the Most relevant spatial-temporal relations. IEEE Access 6:38186–38199

Soysal ÖM (2015) Association rule mining with mostly associated sequential patterns. Expert Syst Appl 42(5):2582–2592

Sulemana I (2012) Assessing over-aged Car legislation as an environmental policy law in Ghana. Int J Bus Soc Sci 3(20)

Sullivan TJ, Driscoll CT, Beier CM, Burtraw D, Fernandez IJ, Galloway JN et al (2018) Air pollution success stories in the United States: the value of long-term observations. Environ Sci Policy 84:69–73

Terry WR, Lee JB, Kumar A (1986) Time series analysis in acid rain modeling: evaluation of filling missing values by linear interpolation. Atmos Environ 20:1941–1943

Tian Y, Yao X, Chen L (2019) Analysis of spatial and seasonal distributions of air pollutants by incorporating urban morphological characteristics. Comput Environ Urban Syst 75:35–48

Villar A, Zarrabeitia MT, Fdez-Arroyabe P, Santurtún A (2018) Integrating and analyzing medical and environmental data using ETL and business intelligence tools. Int J Biometeorol 62(6):1085–1095

Wamba SF, Akter S, Edwards A, Chopin G, Gnanzou D (2015) How 'big data' can make big impact: findings from a systematic review and a longitudinal case study. Int J Prod Econ 165:234–246

Wang D, Wei S, Luo H, Yue C, Grunder O (2017a) A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. Sci Total Environ 580:719–733

Wang H, Zhao L (2018) A joint prevention and control mechanism for air pollution in the Beijing-Tianjin-Hebei region in China based on long-term and massive data mining of pollutant concentration. Atmos Environ 174: 25–42

Wang J, Song G (2018) A deep spatial-temporal ensemble model for air quality prediction. Neurocomputing 314: 198–206

Wang J, Zhang X, Guo Z, Lu H (2017b) Developing an early-warning system for air quality prediction and assessment of cities in China. Expert Syst Appl 84:102–116

Wang L, Zhong B, Vardoulakis S, Zhang F, Pilot E, Li Y et al (2016) Air quality strategies on public health and health equity in Europe—a systematic review. Int J Environ Res Public Health 13(12):1196

Wang S, Paul MJ, Dredze M (2015) Social media as a sensor of air quality and public response in China. J Med Internet Res 17(3)

Westerlund J, Urbain JP, Bonilla J (2014) Application of air quality combination forecasting to Bogota. Atmos Environ 89:22–28

Witten IH, Frank E, Hall MA, Pal CJ (2016) Data mining: practical machine learning tools and techniques. Morgan Kaufmann

World Health Organization (2016). Ambient air pollution: a global assessment of exposure and burden of disease

Wu Y, Zhang F, Shi Y, Pilot E, Lin L, Fu Y et al (2016) Spatiotemporal characteristics and health effects of air pollutants in Shenzhen. Atmos Pollut Res 7(1):58–65

Xie Y, Zhao L, Xue J, Gao HO, Li H, Jiang R et al (2018) Methods for defining the scopes and priorities for joint prevention and control of air pollution regions based on data-mining technologies. J Clean Prod 185:912–921

Xu Y, Yang W, Wang J (2017) Air quality early-warning system for cities in China. Atmos Environ 148:239–257

Yang F, Tan J, Zhao Q, Du Z, He K, Ma Y et al (2011) Characteristics of PM2.5 speciation in representative megacities and across China. Atmos Chem Phys 11(11):5207–5219

Yang G, Huang J, Li X (2018b) Mining sequential patterns of PM2. 5 pollution in three zones in China. J Clean Prod 170:388–398

Yang L, Xu H, Jin Z (2018a). Estimating spatial variability of ground-level PM2.5 based on a satellite-derived aerosol optical depth product: Fuzhou, China

Yang X, Zheng Y, Geng G, Liu H, Man H, Lv Z, He K, de Hoogh K (2017) Development of PM2.5 and NO_2 models in a LUR framework incorporating satellite remote sensing and air quality model data in Pearl River Delta region, China. Environmental Pollution 226:143–153

Yeganeh B, Hewson MG, Clifford S, Knibbs LD, Morawska L (2017) A satellite-based model for estimating PM2.5 concentration in a sparsely populated environment using soft computing techniques. Environ Model Softw 88:84–92

Zhang C, Ni Z, Ni L (2015) Multifractal detrended cross-correlation analysis between PM2.5 and meteorological factors. Physica A: Statist Mech Appl 438:114–123

Zhang NN, Ma F, Qin CB, Li YF (2018) Spatiotemporal trends in PM2.5 levels from 2013 to 2017 and regional demarcations for joint prevention and control of atmospheric pollution in China. Chemosphere 210:1176–1184

Zhang Y, Bocquet M, Mallet V, Seigneur C, Baklanov A (2012) Real-time air quality forecasting. Part I: History, techniques, and current status. Atmos Environ 60:632–655

Zhao C, Song G (2017) Application of data mining to the analysis of meteorological data for air quality prediction: a case study in Shenyang. IOP Conf Ser: Earth Environ Sci 81(1)

Zotteri G, Kalchschmidt M, Caniato F (2005) The impact of aggregation level on forecasting performance. Int J Prod Econ 93:479–491

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

Affiliations

Natacha Soledad Represa[1,2] • **Alfonso Fernández-Sarría**[2] • **Andrés Porta**[1] • **Jesús Palomar-Vázquez**[2]

Alfonso Fernández-Sarría
afernan@cgf.upv.es

Andrés Porta
aporta@quimica.unlp.edu.ar

Jesús Palomar-Vázquez
jpalomav@upvnet.upv.es

[1] Centro de Investigaciones del Medioambiente, National University of La Plata (UNLP), 47 y 115, s/n., B1900AJL La Plata, Argentina

[2] Geo-Environmental Cartography and Remote Sensing Group, Polytechnic University of Valencia, Camí de Vera, s/n. CP, 46022 Valencia, Spain