*Journal of Universal Computer Science, vol. 24, no. 3 (2018), 261-276*
*submitted: 15/10/17, accepted: 5/3/18, appeared: 28/3/18 © J.UCS*

# Air-Pollution Prediction in Smart Cities through Machine Learning Methods: A Case of Study in Murcia, Spain

**Raquel Martínez-España, Andrés Bueno-Crespo, Isabel Timón, Jesús Soto, Andrés Muñoz, José M. Cecilia**
(Department of Computer Engineering, Universidad Católica de Murcia, Spain
{rmartinez, abueno, imtimon, jsoto, amunoz, jmcecilia}@ucam.edu)

**Abstract:** Air-pollution is one of the main threats for developed societies. According to the World Health Organization (WHO), pollution is the main cause of deaths among children aged under five. Smart cities are called to play a decisive role to improve such pollution by first collecting, in real-time, different parameters such as $SO_2$, $NO_x$, $O_3$, $NH_3$, $CO$, $PM_10$, just to mention a few, and then performing the subsequent data analysis and prediction. However, some machine learning techniques may be more well-suited than others to predict pollution-like variables. In this paper several machine learning methods are analyzed to predict the ozone level ($O_3$) in the Region of Murcia (Spain). $O_3$ is one of the main hazards to health when it reaches certain levels. Indeed, having accurate air-quality prediction models is a previous step to take mitigation activities that may benefit people with respiratory disease like Asthma, Bronchitis or Pneumonia in intelligent cities. Moreover, here it is identified the most-significant variables to monitor the air-quality in cities. Our results indicate an adjustment for the proposed $O_3$ prediction models from 90% and a root mean square error less than 11 $\mu/m^3$ for the cities of the Region of Murcia involved in the study.

**Key Words:** Air-pollution monitoring, Ozone, Smart cities, Machine learning, Random forest, Hierarchical clustering.
**Category:** I.2, H.3.5, H.4

## 1 Introduction

Smart cities have become an endless source of urban data. These data range from traffic events to data related to the management of public resources, through indicators about the citizens' quality of life. Among the latter, one of the most important indicators is related to air quality. According to the World Health Organization (WHO), the air pollution is a leading cause of chronic or non-communicable diseases (NCDs), causing over one-third of deaths from stroke, lung cancer and chronic respiratory disease, and one-quarter of deaths from ischaemic heart disease [WHO, 2018]. In fact, this issue is included by the European Union as one of the challenges for smart cities in its H2020 programme, recently debated in the European Forum on Eco-Innovation[1].

Air quality is affected by several factors including airborne particulate matter ($PM$), sulfur dioxide ($SO_2$), Nitrogen dioxide ($NO_2$) and Ozone ($O_3$), just to

[1] http://ec.europa.eu/environment/ecoinnovation2018/1st_forum/case-studies_02_en.html

mention a few. Several works have shown that short-term $O_3$ exposures within a period of 1 to 2 days may be directly related to acute coronary events in middle-aged adults without heart disease [Ruidavets et al., 2005, Brook et al., 2002]. Indeed, the continuous monitoring of these variables can provide firm foundation for creating models to predict hypothetical high-level concentrations of a determinate polluted factor. Several works have been done in developing IoT infrastructures for air pollution monitoring [Al-Ali et al., 2010, Shaikh et al., 2017, Afshar-Mohajer et al., 2018]. However, the fast-growing, tremendous amount of data, collected and stored in large and numerous data repositories, have far exceeded our human ability for comprehension without computational tools. Some efforts have been made to develop expert system and knowledge-based technologies, which typically rely on users or domain experts to *manually* input knowledge into knowledge databases. However, the manual knowledge input procedure is prone to biases and errors and it is extremely costly and time-consuming. The widening gap between data and information calls for the systematic development of *data mining tools*.

In the data mining area, the data is stored electronically and the analysis is automated, or at least augmented, by computers. Some efforts have been done with the idea that patterns in data can be automatically sought, identified, validated, and used for prediction. Data mining is defined in [Witten et al., 2016] as the process of discovering patterns in data and, in [Hand, 2007], it is stated as the discovery of interesting, unexpected or valuable structures in large datasets. Indeed, the process must be automatic or, more usually, semiautomatic and the patterns discovered must be meaningful in a practical sense. Data mining is composed of several techniques, including machine learning and statistics analysis, just to mention few of them. However, the data mining main goal is to target a specific scenario which is modeled with a particular data set in order to deal with a specific problem or situation.

In practice, most tasks that require intelligence also require an ability to induce new knowledge from experiences. A computer program is said to learn some task from experience if its performance at the task improves with experience, according to some performance measure. *Machine learning* investigates how computers can learn, or improve their performance, based on data. A main research area for computer programs is to automatically learn to recognize complex patterns and make intelligent decisions based on data. In [Han et al., 2011] machine learning is classified in four categories: supervised learning, unsupervised learning, semi-supervised learning and active learning.

One of the aims of smart cities is to act on the basis of data obtained through sensors. However, it must be noted that the sensors may cause failures and errors when obtaining data, hence it is necessary to develop a model to predict values of interest in order to control air quality. The main goal of this paper is to analyze different machine learning techniques to predict ozone levels. The aim of this analysis is to obtain the best model for predicting ozone. Thus, in the event of a sensor failure, the model can predict with the least possible error the amount of ozone in the air in order to create an alert if the recommended thresholds are exceeded and take the appropriate measures. Specifically, this analysis has been performed in four cities at the Region of Murcia (Spain) taking real data from 4 stations for air quality measurement.

The rest of the paper is structured as follows. Section 2 describes work related to air quality and ozone prediction in cities. Section 3 presents the machine learning techniques used for ozone prediction as well as for the study of the most important variables to consider for such prediction. Section 4 shows the assessment performed to obtain the best models for the ozone prediction and a clustering of stations according to the similarity of the data. Finally, Section 5 describes the conclusion and future work of this paper.

## 2 Related Work

Smart cities have attracted considerable attention in the context of urban development policies. The Internet and broadband networking technologies are seen as enablers of e-services and are becoming increasingly important for urban development, while cities increasingly assume a key role as drivers of innovation in areas such as health, inclusion, environment and business [Schaffers et al., 2011]. Thus, for example, there are many studies related to traffic control in cities [Bui and Jung, 2017, Shaghaghi et al., 2017], and this topic is related to the environment in smart cities. In this same area and related to traffic control is the topic of air quality monitoring. Many of the works in the area of air pollution in smart cities focus on the monitoring of parameters considered as pollutants. Thus, in [Al-Ali et al., 2010] a GPRS-Sensor array system is proposed to report real-time pollution level in a Google map. Basically, this system is empowered with $CO$, $NO_2$ and $SO_2$ pollution sensors augmented with GPS data including location, date and time. Authors show a proof of concept for the city of Sharjah in the United Arab Emirates. A more advanced approach is presented in [Jin et al., 2014], where an IoT-based infrastructure called IDRA offers several environmental monitoring services. Among them, authors highlight the vigilance of parameters such as hydrocarbons and oxides of nitrogen. Due to the rapid evolution of Information Technology (IT), we have entered the age of Big Data in multiple areas of research (see for example [Jung, 2017a, Jung, 2017b]). A Big Data analytics-based approach [Rathore et al., 2016] uses ozone, $CO$, $NO_2$ and $SO_2$ levels to, along with data from smart homes, traffic, time, surveillance, etc., assist in urban planning decision making. However, these works do not propose any technique for predicting pollution for the next days or identifying related factors.

With an ever-increasing air pollution ratios, it is necessary to implement efficient air quality monitoring models, obtained from the data collected by pollution sensors, that help to predict the concentration of air pollutants and provide assessment of air pollution in each area. Hence, air quality evaluation and prediction have become an important research area. In relation to works in the literature that do take into account the prediction of air pollution, there is a clear majority of use of artificial neural networks (ANN) compared to other models such as multiple linear regression(MLR) [Kurt et al., 2008, Azid et al., 2013, Ahmad et al., 2017, Azid et al., 2017]. Being this the general trend, it is important to observe that ANNs also present some weakness for this topic, as identified by [Zhang and Ding, 2017]: They have poor generalization, falling in local minimum with relative ease; they do not have an analytical method for model selection; and they follow a long-running process to obtain the most accurate model. Finally, it is worth mentioning new approaches that combine machine learning techniques with the use of social media data to understand and improve predictions on air pollution [Ravi et al., 2017].

In [Kang et al., 2018] various big-data and machine learning based techniques for air quality forecasting are investigated. The paper reviews the published research results relating to air quality evaluation using methods of artificial intelligence, decision trees, deep learning, etc. On the other hand, in [Reid et al., 2017] is explored the implementation of an Internet of Things multiagent system distributed along the roadway to collect and share vehicular data among its nodes and then process the data using a machine learning algorithm for inference of vehicle types. In [Li et al., 2017] is proposed a deep learning model to estimate air pollution throughout the city, utilizing the readily available urban data as proxy data. As with many big data driven approaches, the proxy data may be sparse/missing. The authors propose the M-BP algorithm to recover/fill in such missing data. The potential usage of machine learning and reduced-order modeling techniques to mitigate some of these limitations is discussed in [Keller et al., 2017], where the authors find that this approach shows promising initial results for important air pollutants such as Ozone, predicting concentrations that deviate less than 10% from the values computed by the traditional model.

The specific problem of the prediction of Ozone and PM10 is addressed in [Corani, 2005], using to this end several statistical approaches. In particular, they use feed-forward neural networks (FFNNs), which have been extensively used for the prediction of air quality, and they are compared to two different machine learning approaches: lazy learning and pruned neural networks.

Application of a novel classifier ($\sigma - FLNMAP$) [Athanasiadis et al., 2003] is introduced for estimating the ozone concentration level in the atmosphere. The $\sigma - FLNMAP$ classifier gets better results (with only a few rules) compared to

FFNNN and C4.5 algorithm.

## 3 Machine learning methods for air-pollution monitoring

In this work several automatic learning techniques are evaluated to predict the level of ozone in smart cities. For this purpose, several techniques have been selected taking as criteria the interpretability of the models they obtain. The techniques used have been extensively tested and are capable of providing good performance. The techniques used are: Bagging, Random Committee, Random Forest, a decision tree and an instance-based technique. A summary of the basic fundamentals of the techniques is explained below:

- **Bagging**

  Bagging is a multi classifier that learns several classifiers and output is a composition of the result that each of them [Breiman, 1996]. The base classifier can be based on different techniques, for example, trees, rules, instances, etc. In this case, the classifier used in this paper is REPTree. REPTree builds a decision/regression tree using information gain and prunes it using reduced error pruning (with adjustment). This tree classifies the values of the numerical attributes only once. This tree has a behavior similar to the decision tree C4.5 [Quinlan, 2014].

- **Random Committe**

  Random Committe is an ensemble of random base classifiers. Each base classifier is constructed using a different random number of seeds (but using the same data). The final prediction is a direct average of the predictions generated by the individual base classifiers. The base classifier used in this paper is the Random Tree. Random Tree [Kalmegh, 2015] is a decision tree that considers K randomly selected attributes at each node. It does not prune. It also has an option to estimate the target mean for regression based on a hold-out set (backfitting).

- **Random Forest**

  Random forest [Breiman, 2001] is an ensemble composed of decision trees where each tree depends on the values of a random vector sampled independently and with the same distribution for all the trees in the forest. The error for the forest tends to stabilize from a certain elevated number of trees. This is because the error depends on the quality of each individual tree and the correlation between those trees.

- **Decision Tree**

The decision tree used in this case is the M5P [Wang and Witten, 1997]. This tree is an improved version of [Quinlan et al., 1992]. The basic idea for building this tree model is quite straightforward. In a first stage an induction decision tree is constructed where instead of maximizing the gain of information within each node a division criteria is used to minimize the intra-subset variation. In the second stage, a pruning is carried out inside the nodes, replacing the node if necessary with a regression plane.

- **k Nearest Neighbors (kNN)**

  The technique (kNN) [Aha et al., 1991] is a type of instance-based learning, or lazy learning, in which the function is only approached locally and all calculations are postponed until classification. This technique is used in both classification and regression, it has no training phase and it calculates the nearest neighbours to a given instance using distance or similarity functions. For regression the output consist of the average of the values of its k nearest neighbors.

- **Hierarchical cluster**

  Hierarchical clustering technique [Langfelder et al., 2007] defines the cluster distance between two clusters to be the maximum distance between their individual components. At every stage of the clustering process, the two nearest clusters are merged into a new cluster. In this case, the technique is based on the Discrete Wavelet Transform (DWT) [Wickerhauser, 1996], which is a useful feature extraction technique often used to measure dissimilarity between Time Series. DWT performs a scale-wise decomposing of the time series in such a way that most of the energy of the time series can be represented by only a few coefficients. The basic idea is to replace the original series by their wavelet approximation coefficients in an appropriate scale, and then to measure the dissimilarity between both.

## 4 Evaluation

The techniques explained in section 3 are now evaluated by means of a series of datasets obtained from four air-quality measurement stations in four different cities at the Region of Murcia, Spain. The prediction is performed on $O_3$, being one of the most polluting agents in the air. The assessment is divided into two parts. Firstly, the $O_3$ prediction results are calculated for the different techniques explained in section 3. The Weka tool has been used for this evaluation [Hall et al., 2009]. Secondly, a hierarchical clustering is carried out to evaluate how many models would be necessary to predict ozone in the Region of Murcia. To obtain the endogram through the hierarchical cluster the R language is used, specifically the TSclust package [Manso et al., 2017]. For the two experimental approaches the datasets are described next.

### 4.1 Dataset

The data used in this experiment are real data obtained from four atmospheric stations [2] in the Region of Murcia. These stations are located in the cities of Alcantarilla, Aljorra, Lorca and Caravaca. The data used covers the average per hour of different climatic parameters and chemical elements affecting air quality each day for the years 2013 and 2014 for each station. Not all stations have the same measuring instruments and therefore not all stations analyze all elements affecting air quality. Moreover, as measuring instruments do not always work properly, when the dataset contained missing data in any of the input variables used, the instance has been discarded, so each dataset contains a different number of instances.

Table 1 shows the description of the datasets used to predict ozone in the aforementioned four cities of the Region of Murcia. In this table, the "N.Inst." column indicates the number of instances that each dataset contains whereas the "Inputs" column indicates the variables that are taken into account in the datasets to predict ozone. For the "Alcantarilla" station, new air quality sensors were added in 2014, hence for 2013 there are 10 input variables whereas for 2014 there are 13 input variables. The datasets called "Alcantarilla2" consist in the fact that for 2014 those new variables have been eliminated and the same input variables have been left as in 2013 in order to establish if those new variables incorporated in 2014 are significant or not.

The inputs used to predict Ozone, shown in an abbreviated form in Table 1, are measured in microgram per cubic meter ($\mu/m^3$) and consist of Nitrogen Monoxide (NO), Nitrogen Dioxide ($NO_2$), Sulfur Dioxide($SO_2$), Total Nitrogen Oxides (NOX), Particulate matter in suspension $< 10\mu\ g(PM_{10})$, Benzeno ($C_6H_6$), Toluene ($C_7H_8$) and Xileno (XIL). The rest of the elements are Temperature (TMP) measured in degrees Celsius ( $^oC$), Relative Humidity (HR) measured in %, wind direction (DD) in grades, Wind speed (VV) in meters per second (m/s), Atmospheric pressure (PRB) in bar and Solar Radiation (RS) in watts per square meter(w/$m^2$). Finally, the variable to be predicted $O_3$ is also measured in $\mu/m^3$.

### 4.2 Parameters

The machine learning techniques have been validated using different parameters where the best results are shown in Table 2. For this table it should be clarified

[2] https://sinqlair.carm.es/calidadaire/Default.aspx

**Table 1:** Description of the datasets for ozone prediction.

| | 2013 | | 2014 | |
|---|---|---|---|---|
| Datasets | N.Inst. | Inputs | N.Inst. | Inputs |
| Alcantarilla | 8496 | NO, $NO_2$, $SO_2$, TMP, HR, NOX, DD, VV, $PM_{10}$ | 8496 | NO, $NO_2$, $SO_2$, TMP, HR, NOX, DD, VV, $C_6H_6$, $C_7H_8$, XIL, $PM_{10}$ |
| Alcantarilla2 | - | - | 8496 | NO, $NO_2$, $SO_2$, TMP, HR, NOX, DD, VV, $PM_{10}$ |
| Aljorra | 6093 | NO, $NO_2$, $SO_2$, TMP, HR, NOX, DD, PRB, RS, VV, $PM_{10}$ | 8348 | NO, $NO_2$, $SO_2$, TMP, HR, NOX, DD, PRB, RS, VV, $PM_{10}$ |
| Lorca | 7982 | NO, $NO_2$, $SO_2$, TMP, HR, NOX, DD, RS, VV, $PM_{10}$ | 7865 | NO, $NO_2$, $SO_2$, TMP, HR, NOX, DD, RS, VV, $PM_{10}$ |
| Caravaca | 8341 | NO, $NO_2$, TMP, HR, NOX, DD, PRB, RS, VV, $PM_{10}$ | 8492 | NO, $NO_2$, TMP, HR, NOX, DD, PRB, RS, VV, $PM_{10}$ |

that in the Random Committee technique the value of K for Random Tree consists of $log_2 A + 1$, where $A$ is the number of input values. Likewise for the decision tree, the parameter “Min Number Samples” refers to a node being a leaf when it contains that minimum number of examples.

### 4.3 Ozone prediction

In this section, the machine learning techniques proposed for predicting $O_3$ are assessed and analyzed. This assessment is performed by a 3-fold cross validation; i.e. the database is divided into three groups where one group is selected for the evaluation and the other two for training. The three groups are eventually used for evaluation. Moreover, the quality and reliability of our models are measured by the Mean Absolute Error (MAE) and Root Mean Squared Error

**Table 2:** Relevant parameters for the selected machine learning techniques.

| Techniques | Parameters | |
|---|---|---|
| Bagging | Base Classifier: | REPTree |
| | Iterations: | 20 |
| | % Bag Size: | 100% |
| Random Committe | Base Classifier: | Random Tree |
| | K Value for Random Tree: | $log_2 A + 1$ |
| | Iterations: | 20 |
| Random Forest | Base Classifier: | C4.5 |
| | Number Trees: | 150 |
| | Minimum Features: | 1 |
| Decision Tree | Min Number Samples: | 4 |
| KNN | K value: | 2 |
| Hierarchical cluster | Distance: | DWT<br>Euclidean |

(RMSE). Finally, the robustness and suitability are evaluated through the determination coefficient $R^2$. The measurements MAE, RMSE and $R^2$ are defined in the equations 1, 2 and 3, respectively.

$$MAE = \frac{\sum_{i=1}^{n} |v_p - v_r|}{n} \tag{1}$$

where $n$ is the number of instances, $v_p$ is the value predicted by the model and $v_r$ is the actual ozone value.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_p - v_r)^2} \tag{2}$$

$$R^2 = \frac{\sigma^2_{v_p,v_r}}{\sigma^2_{v_p}\sigma^2_{v_r}} \tag{3}$$

where $\sigma_{v_p,v_r}$ is the covariance of $v_p$, $v_r$ and $\sigma_{v_p}$ and $\sigma_{v_r}$ is the standard deviation of the variable $v_p$ and $v_r$ respectively.

Table 3 shows the error metrics of the proposed machine learning techniques in terms of RMSE and MAE in the cities involved in the study. It is noteworthy to highlight that the RMSE and MAE of the Alcantarilla2 dataset ("Alcant.2") is relatively similar to the Alcantarilla ("Alcant.") dataset for the year 2014. This means the new sensors (i.e. variables) introduced at Alcantarilla station during 2014 do not have a great influence on achieving a better ozone prediction.

Furthermore, the random forest algorithm obtains, in general, a lower RMSE and MAE than the other machine learning strategies. The Random Committee

Table 3: RMSE and MAE obtained by machine learning techniques in the prediction of $O_3$ for the years 2013 and 2014.

| Techniques | | Bagging | | Random Committee | | Random Forest | | M5P Tree | | KNN | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Data | Years | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| Alc | 2013 | 8.31 | 10.99 | 8.01 | 10.83 | 7.65 | 10.20 | 8.48 | 11.19 | 8.87 | 12.33 |
| | 2014 | 8.03 | 10.65 | 7.66 | 10.26 | 7.33 | 9.77 | 8.66 | 11.47 | 9.43 | 13.07 |
| Alc2 | 2014 | 8.22 | 10.97 | 8.00 | 10.78 | 7.68 | 10.24 | 8.58 | 11.41 | 8.87 | 12.33 |
| Alj | 2013 | 9.17 | 12.14 | 8.62 | 11.68 | 8.34 | 11.11 | 8.57 | 11.36 | 9.32 | 12.98 |
| | 2014 | 7.63 | 9.79 | 7.35 | 9.59 | 7.10 | 9.16 | 8.06 | 10.29 | 7.70 | 10.37 |
| Lorca | 2013 | 8.95 | 11.72 | 8.61 | 11.50 | 8.25 | 10.89 | 9.46 | 12.31 | 10.04 | 13.64 |
| | 2014 | 8.53 | 11.09 | 8.13 | 10.77 | 7.87 | 10.30 | 9.25 | 11.92 | 9.05 | 12.36 |
| Car | 2013 | 8.50 | 11.07 | 8.13 | 10.80 | 7.90 | 10.35 | 9.19 | 11.94 | 8.93 | 12.33 |
| | 2014 | 8.54 | 11.12 | 7.96 | 10.68 | 7.77 | 10.29 | 9.30 | 12.06 | 9.71 | 13.27 |

is placed in second position according with the error ratios. The overall comparison of the targeted machine learning techniques is shown in Figure 1, where Figure 1(a) refers to 2013 and Figure 1(b) to 2014. These two figures shows several relevant points. In 2013, the city with the least error is Caravaca followed by Aljorra. In 2014, the city with the best prediction was Aljorra followed by Alcantarilla, the random forest predicted. The worst performance machine learning technique studied (i.e. having the highest RMSE and MAE) was KNN. This fact happened using all datasets for the two years studied, so this technique obtains unsatisfactory results to predict $O_3$ and we would not advise the use of this technique to predict $O_3$ whenever this sensor may fail in a smart city environment.

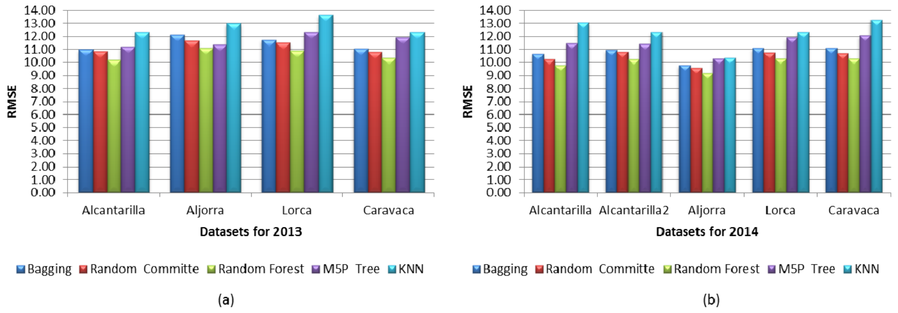

Figure 1: Comparison of the RMSE for the years 2013 (a) and 2014 (b) of the different machine learning techniques.

Table 4 shows the determination coefficient $R^2$ to assess the quality of the results for our models. Above $0.75R_2$ could be considered as satisfactory (let us remind the reader that $R_2 = 1$ means a perfect fit). Therefore, Table 4 shows that most of the models generally obtain a good fit and therefore they obtain satisfactory and reliable results. However, it should be underlined that for 2013 and for the city of Caravaca the adjustment of the different models is worse.

Table 4: Adjustment of the models obtained by machine learning techniques for the years 2013 and 2014.

| Techniques | | Bagging | Random Committee | Random Forest | M5P Tree | KNN |
|---|---|---|---|---|---|---|
| Datasets | Years | $R^2$ | $R^2$ | $R^2$ | $R^2$ | $R^2$ |
| Alcantarilla | 2013 | 0.895 | 0.898 | 0.910 | 0.891 | 0.870 |
| | 2014 | 0.908 | 0.911 | 0.920 | 0.900 | 0.870 |
| Alcantarilla2 | 2014 | 0.913 | 0.919 | 0.927 | 0.899 | 0.871 |
| Aljorra | 2013 | 0.792 | 0.807 | 0.826 | 0.897 | 0.766 |
| | 2014 | 0.801 | 0.808 | 0.826 | 0.780 | 0.779 |
| Lorca | 2013 | 0.832 | 0.839 | 0.856 | 0.815 | 0.780 |
| | 2014 | 0.849 | 0.858 | 0.870 | 0.835 | 0.817 |
| Caravaca | 2013 | 0.679 | 0.695 | 0.722 | 0.626 | 0.616 |
| | 2014 | 0.742 | 0.761 | 0.780 | 0.752 | 0.645 |

Figure 2 shows that the prediction for the Caravaca city is always lower on average than the other cities. Moreover, although the best model studied is random forest, for the year 2013 and for the city of Aljorra, the technique with the best suitability is the M5P decision tree. Regarding the prediction error, the technique with the worst suitability is KNN.

Although Figures 1 and 2 state the random forest technique obtains the best models to predict $O_3$, a non-parametric statistical test has been performed to validate this statement; the Wilcoxon Signed Ratings Test is used [Kruskal, 1957]. This test compares two paired groups and thus it can be used to test when the null hypothesis indicates that two populations have the same continuous distribution. The Wilcoxon test confirms that the Random Forest technique provides better results and a better fit than the other machine learning techniques with 99% confidence level. The second and third techniques with a better fit for the $O_3$ prediction are Random Committee and Bagging respectively.

Finally, it is noteworthy to highlight that the techniques used in this analysis enables the result interpretation in a simple manner. Therefore, we are going to use decision tree techniques to analyze which variables are the most important

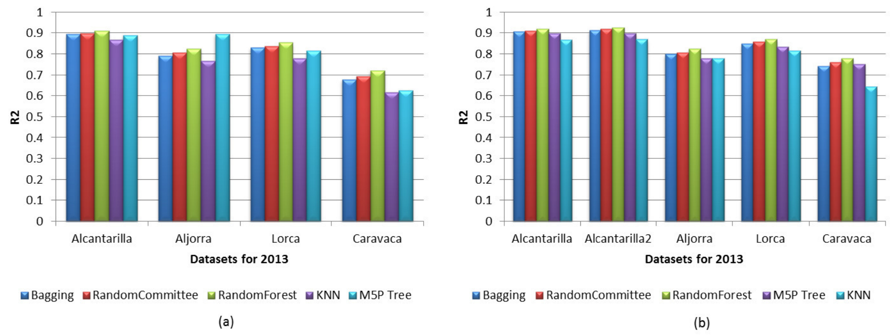

Figure 2: Suitability of the models according to $R^2$ for the cities of the Region of Murcia for the years 2013 (a) and 2014 (b).

in predicting $O_3$. Thus, and taking into account the variables that are measured in all cities, the most influential variables to predict $O_3$ are: $NO_X$, TMP, DD, VV, HR, $SO_2$, NO and $PM_{10}$.

### 4.4 Clustering cities according to their $O_3$

Once the best model for predicting ozone has been selected, a hierarchical clustering technique is applied to identify how many models would be needed to predict the ozone in the Region of Murcia. This analysis has been carried out for the year 2014. Two different measures have been used for this purpose: on the one hand, DWT (Dissimilarity for Time Series Based on Wavelet Feature Extraction) and on the other hand, Euclidean distance measurement. The reason for using DWT is to analyze the results obtained by treating the air quality data collected as time series.

Finally, Figure 3 shows the endogram obtained with the 2014-data for the four cities studied in this paper. The endogram shows similar results whenever the euclidean distance or the DWT distance are used. These results show that it is not necessary to deal with the data as a time series. The endogram also unifies the cities of Lorca, Alcantarilla and Aljorra, putting off the city of Caravaca. Caravaca has different air quality ratio than the other 3 cities and therefore it needs a specific model in order to provide an accurate ozone prediction. Thus, by placing the cities on the map of the Region of Murcia, the behaviour of air quality in Murcia is referenced as shown in Figure 4.

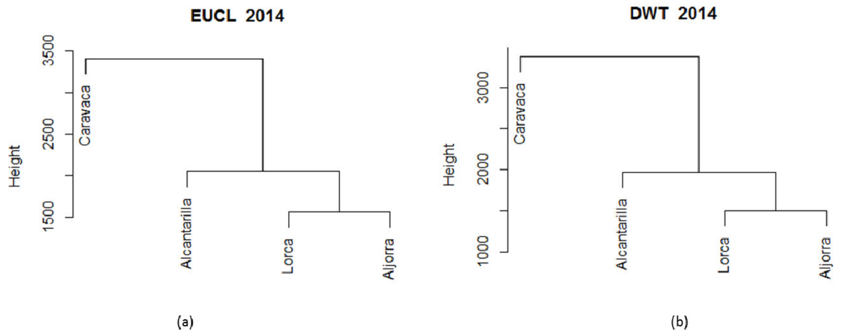

Figure 3: Endogram using the two distances for hierarchical clustering: (a) Using the Euclidean distance; (b) Using DWT distance.

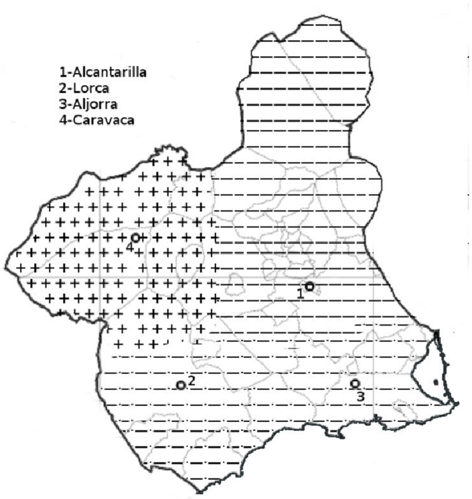

Figure 4: Map of the Region of Murcia by air quality zones classified by the clustering technique.

## 5 Conclusions and Future Work

We are witnessing the steady transaction to smart cities where machine learning is called to play a decisive role. Among the main issues that are currently worth of attention in developed countries is the pollution concentration in city areas causing chronic diseases. Indeed, the and analysis of pollutants such as $SO_2$, $NO_x$, $O_3$, $NH_3$, $CO$ and $PM_10$ through machine learning techniques may help to predict pollution peaks as well as to establish thresholds and action plans for local authorities, drivers, industries, etc. In this paper, we analyze different machine learning techniques to predict the $O_3$ levels, one of the more harmful air-pollution parameter. The machine learning techniques studied in this work have been Random Forest, Decision Tree, Random Committee, Bagging and KNN. The technique that obtains the best fit in general is Random Forest, being this assertion validated by statistical tests. The results indicate an $R^2$ setting between 80% and 90% overall and an $O_3$ prediction error less than 11 $\mu/m^3$. It is also important to note that among the parameters that most influence the ozone prediction we have found climatic variables related to temperature, humidity and wind. In addition, hierarchical clustering indicates that the air-pollution monitoring areas in the Region of Murcia can be divided into two zones only so as to create two general $O_3$ prediction models for the entire Region. These two areas would be the cities of Lorca, Alcantarilla and Aljorra on one side and Caravaca on the other.

As future work, new parameters such as $PM_{10}$ and $SO_2$, that seriously affect air quality as well, must be analyzed and studied to create models that help to predict them. Another extension will be the automatic generation of recommendations to local authorities, drivers and other related actors in the influence of air-quality factors according to the alerts raised by our system.

## Acknowledgements

This work is supported by the Spanish MINECO under grant TIN2016-78799-P (AEI/FEDER, UE).

## References

[Afshar-Mohajer et al., 2018] Afshar-Mohajer, N., Zuidema, C., Sousan, S., Hallett, L., Tatum, M., Rule, A. M., Thomas, G., Peters, T., and Koehler, K. (2018). Evaluation of low-cost electro-chemical sensors for environmental monitoring of ozone, nitrogen dioxide and carbon monoxide. Journal of occupational and environmental hygiene.

[Aha et al., 1991] Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. Machine learning, 6(1):37–66.

[Ahmad et al., 2017] Ahmad, Z., Rahim, N. A., Bahadori, A., and Zhang, J. (2017). Air polluiton index prediction using multiple neural networks. IIUM Engineering Journal, 18(1):1–12.

[Al-Ali et al., 2010] Al-Ali, A. R., Zualkernan, I., and Aloul, F. (2010). A mobile gprs-sensors array for air pollution monitoring. IEEE Sensors Journal, 10(10):1666–1671.

[Athanasiadis et al., 2003] Athanasiadis, I. N., Kaburlasos, V. G., Mitkas, P. A., and Petridis, V. (2003). Applying machine learning techniques on air quality data for real-time decision support. In First international NAISO symposium on information technologies in environmental engineering (ITEE'2003), Gdansk, Poland.

[Azid et al., 2013] Azid, A., Juahir, H., Latif, M. T., Zain, S. M., and Osman, M. R. (2013). Feed-forward artificial neural network model for air pollutant index prediction in the southern region of peninsular malaysia. Journal of Environmental Protection, 4(12):1.

[Azid et al., 2017] Azid, A., Rani, N., Samsudin, M., Khalit, S., Gasim, M., Kamarudin, M., Yunus, K., Saudi, A., and Yusof, K. (2017). Air quality modelling using chemometric techniques. Journal of Fundamental and Applied Sciences, 9(2S):443–466.

[Bello-Orgaz et al., 2016] Bello-Orgaz, G., Jung, J. J., and Camacho, D. (2016). Social big data: Recent achievements and new challenges. Information Fusion, 28:45-59.

[Breiman, 1996] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2):123–140.

[Breiman, 2001] Breiman, L. (2001). Random forests. Machine learning, 45(1):5–32.

[Brook et al., 2002] Brook, R. D., Brook, J. R., Urch, B., Vincent, R., Rajagopalan, S., and Silverman, F. (2002). Inhalation of fine particulate air pollution and ozone causes acute arterial vasoconstriction in healthy adults. Circulation, 105(13):1534–1536.

[Bui and Jung, 2017] Bui, K.-H. N. and Jung, J. J. (2017). Internet of agents framework for connected vehicles: A case study on distributed traffic control system. Journal of Parallel and Distributed Computing 116:89-95.

[Corani, 2005] Corani, G. (2005). Air quality prediction in milan: feed-forward neural networks, pruned neural networks and lazy learning. Ecological Modelling, 185(2-4):513–529.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. ACM SIGKDD explorations newsletter, 11(1):10–18.

[Han et al., 2011] Han, J., Pei, J., and Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[Hand, 2007] Hand, D. J. (2007). Principles of data mining. Drug safety, 30(7):621–622.

[Hong and Jung, 2018] Hong, M. and Jung, J. J. (2018). Multi-Sided recommendation based on social tensor factorization. Information Sciences, 447:140-156.

[Jin et al., 2014] Jin, J., Gubbi, J., Marusic, S., and Palaniswami, M. (2014). An information framework for creating a smart city through internet of things. IEEE Internet of Things Journal, 1(2):112–121.

[Jung, 2017a] Jung, J. E. (2017b). Discovering Social Bursts by Using Link Analytics on Large-Scale Social Networks. Mobile Networks & Applications, 22(4):625–633.

[Jung, 2017b] Jung, J. J. (2017a). Computational collective intelligence with big data: Challenges and opportunities. Future Generation Computer Systems, 66:87–88.

[Kalmegh, 2015] Kalmegh, S. (2015). Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. Int. J. Innov. Sci. Eng. Technol, 2(2):438–446.

[Kang et al., 2018] Kang, G. K., Gao, J. Z., Chiao, S., Lu, S., and Xie, G. (2018). Air quality prediction: Big data and machine learning approaches. International Journal of Environmental Science and Development, 9(1):8–16.

[Keller et al., 2017] Keller, C. A., Evans, M. J., Kutz, J. N., and Pawson, S. (2017). Machine learning and air quality modeling. In Big Data (Big Data), 2017 IEEE International Conference on, pages 4570–4576. IEEE.

[Kruskal, 1957] Kruskal, W. H. (1957). Historical notes on the wilcoxon unpaired two-sample testhistorical notes on the wilcoxon unpaired two-sample test. Journal of the American Statistical Association, 52(279):356–360.

[Kurt et al., 2008] Kurt, A., Gulbagci, B., Karaca, F., and Alagha, O. (2008). An online air pollution forecasting system using neural networks. Environment International, 34(5):592–598.

[Langfelder et al., 2007] Langfelder, P., Zhang, B., and Horvath, S. (2007). Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. Bioinformatics, 24(5):719–720.

[Li et al., 2017] Li, V. O., Lam, J. C., Chen, Y., and Gu, J. (2017). Deep learning model to estimate air pollution using m-bp to fill in missing proxy urban data. In GLOBECOM 2017-2017 IEEE Global Communications Conference, pages 1–6. IEEE.

[Manso et al., 2017] Manso, P. M., Vilar, J. A., and Montero, M. P. (2017). Package 'TScluster'. 62(1):1–43.

[Quinlan, 2014] Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

[Quinlan et al., 1992] Quinlan, J. R. et al. (1992). Learning with continuous classes. In 5th Australian joint conference on artificial intelligence, volume 92, pages 343–348. Singapore.

[Rathore et al., 2016] Rathore, M. M., Ahmad, A., Paul, A., and Rho, S. (2016). Urban planning and building smart cities based on the internet of things using big data analytics. Computer Networks, 101:63–80.

[Ravi et al., 2017] Ravi, N., Manoranjani, R., Seshadri, K., et al. (2017). Leveraging social networks for smart cities: A case-study in mitigation of air pollution. In International Conference on Intelligent Information Technologies, pages 179–193. Springer.

[Reid et al., 2017] Reid, A. R., Pérez, C. R. C., and Rodríguez, D. M. (2017). Inference of vehicular traffic in smart cities using machine learning with the internet of things. International Journal on Interactive Design and Manufacturing (IJIDeM), pages 1–14.

[Ruidavets et al., 2005] Ruidavets, J.-B., Cournot, M., Cassadou, S., Giroux, M., Meybeck, M., and Ferrières, J. (2005). Ozone air pollution is associated with acute myocardial infarction. Circulation, 111(5):563–569.

[Schaffers et al., 2011] Schaffers, H., Komninos, N., Pallot, M., Trousse, B., Nilsson, M., and Oliveira, A. (2011). Smart cities and the future internet: Towards cooperation frameworks for open innovation. In The future internet assembly, pages 431–446. Springer.

[Shaghaghi et al., 2017] Shaghaghi, E., Jabbarpour, M. R., Noor, R. M., Yeo, H., and Jung, J. J. (2017). Adaptive green traffic signal controlling using vehicular communication. Frontiers of Information Technology & Electronic Engineering, 18(3):373–393.

[Shaikh et al., 2017] Shaikh, F. K., Zeadally, S., and Exposito, E. (2017). Enabling technologies for green internet of things. IEEE Systems Journal, 11(2):983–994.

[Wang and Witten, 1997] Wang, Y. and Witten, I. H. (1997). Inducing model trees for continuous classes. In Proceedings of the Ninth European Conference on Machine Learning, pages 128–137.

[WHO, 2018] (2018). World health Organization (WHO) air pollution programme. http://www.who.int/airpollution/en/. Accessed: 2018-03-26.

[Wickerhauser, 1996] Wickerhauser, M. V. (1996). Adapted wavelet analysis: from theory to software. AK Peters/CRC Press.

[Witten et al., 2016] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[Zhang and Ding, 2017] Zhang, J. and Ding, W. (2017). Prediction of air pollutants concentration based on an extreme learning machine: the case of hong kong. International journal of environmental research and public health, 14(2):114.