Ozone Pollution Forecast based on Neural Networks and Decision Trees

Nahun Loya[1], Ivan Olmos[1]

Benemérita Universidad Autónoma de Puebla
Puebla, México
nahun.loya@gmail.com, ivanop_rkl@yahoo.com.mx

Abstract. Nowadays, air pollution is one of the most important problems for the modern society. Many efforts have been conducted with the aim to control and decrease air contaminants, such as chemicals, particulate matter, or biological materials. This paper presents an approach for ozone forecast based on air quality predictors, where machine learning models such as artificial neural networks and decision trees are used. The data considered in this report were recollected between 2010 and 2011 from three different sites of the Atmospheric Monitoring System in Mexico City (SIMAT). Based on our experimental results, it is possible to predict levels of ozone with a high accuracy, and construct rules that can be used for decision support systems.

Keywords: Decision Tree, Air quality forecast, Neural networks, Ozone.

1 Introduction

Big cities such as Los Angeles, Tokyo, Moscow, and Mexico city, are affected by environmental pollution. These cities monitoring the air quality in the troposphere, recording the progress of several air contaminants emitted by vehicles, industries, citizens, and so on[1]. One of the most important pollutants is the ozone, which is a triatomic molecule formed by oxygen atoms. The ozone is considered a powerful oxidant that reacts rapidly with other chemicals [2].

Humans (and other biological organisms) exposed to constant high levels of pollution represent an important health public problem. Many studies report that citizen in Mexico city have a reduction in their lung functions, with severe complications in the respiratory tract [2].

Important cities around the wold implement comprehensive air quality management programs. For example, in Mexico city the government strengthened and began to enforce a vehicle inspection, together with programs such as "no driving day". In addition, air quality is constantly monitored, with the aim to verify if that controls of emission are (or are not) having positive effects.

In this way, in the present work is proposed the development of a tool capable to predict high levels of ozone based on a indirect detection strategy, considering several chemicals and atmospheric variables that different authors have shown to be predictors for the levels of ozone. The variables considered in

this work are: Carbon monoxide (CO), Nitrogen Dioxide (NO_2), Sulfur dioxide (SO_2), Temperature (TMP), Relative humidity (RH), Speed wind (WSP) and Wind direction (WDR). In our study, two strategies of machine learning were considered: artificial neural networks (ANN) and decision trees (DT). In our experiments, the implementation of ANN and DT available in Weka was used [3].

The main objective is to find a good model based on a neural networks and decision trees to predict forecast air quality (with respect to ozone) based on attributes values measured in previous hours. In our experiments, we use a dataset of the Atmospheric Monitoring System (SIMAT, for its acronym in Spanish). First, the data was analyzed with a descriptive statistic for known it. After, data was processing using the KDD process, considering: data cleaning, data integration, and data reduction. After that, a supervised learning process was implemented. In our experiments, ANN's and DT's were used as machine learning algorithms. All our experiments were validated with a cross validation. Based on our results, it is possible to predict high levels of ozone with an accuracy superior of 90%. Moreover, since decision trees are descriptive models, it is possible to build decision rules that can be used in a logical database.

This paper is organized as follows: section 2 shows several studies that have been developed to predict air quality levels using the statistical and other schemes such as neural networks. Section 3, describe the case of study, as well as the locations of weather stations considered in this case of study. Section 4 describes how the integration and cleaning of data is done. Finally section 5 presents the results and models obtained.

2 Related Work

In Mexico, as well as in many countries, there are attempts in many different scientific fields trying to evaluate ozone pollution [4], [5], [6]. As an example, in Mexico city some official institutions and academic works try to predict areas with high levels of pollution, which are dangerous for the citizens.

Some works try to determine levels of ozone by direct measurement recollected in monitoring sites. As results of these efforts, some systems have been implemented [7]. As an example, a monitoring station at the Pedregal in Mexico city was established, that analyze three models for predicting ozone concentrations based on 19 semi-annual average data.

On the other hand, several statistical studies have been conducted to infer and predict air quality with respect to specific pollutants, such as ozone. For example, Seinfeld *et al.*[8] proposed to measure ozone trends with the maximum daily and the average-maximum daily. Another simple way to measure these trends can be determined by the weather conditions associated with those days, where there are high concentrations of ozone. In other words, it is possible to predict high, moderate and low levels of ozone based on the previous observations of the environment.

In the last years, alternative approaches based on machine learning techniques have been proposed with the aim to predict high levels of ozone. For example Aguirre *et al* [9], show the importance of neural networks for forecast ozone: they use a multilayer perceptron model to predict ozone levels for the Pais Vasco in Spain. In a different work, Barai S.V. *et al.* [10] trained neural networks to predict air quality with a limited number of data, that were obtained fromo the U.S. EPA and the Tata Energy Research. According with the reported results, it is possible to predict levels of air pollution with an acceptable accuracy.

Based on the above mentioned, we propose to construct a model based on machine learning algorithms (neural networks and decision trees), capable to predict high levels of ozone in Mexico city, considering different weather variables that will be explained later.

3 Case of Study

In Mexico City, there are many meteorological stations that reports conditions of the weather hourly, including variables related with the air pollution. In this study, dataset comes from of the SIMAT monitoring network, including Pedregal, Tlalnepantla and Xalostoc stations in Mexico city. The dataset was recollected between January 2010 to December 2011.

The dataset include a set of attributes proposed by Seinfeld [8] (see Table 1), which are considered as favorable factors for the presence of high levels of ozone in the air. The first column of Table 1 shown pollutants and atmospheric variables (with their abbreviations), and the second column shown the highest values that are admitted by the Official Mexican Norm per each pollutant. Finally, the units of measure of each atmospheric or chemical variable is shown in the last column [2].

Pollutant/Atmosferic variable	Value of the Standard	Unit of Measure
Ozone (O_3)	0.11 ppm	Parts per million (ppm)
Carbon monoxide (CO)	11 ppm	Parts per million (ppm)
Nitrogen Dioxide (NO2)	0.21 ppm	Parts per million (ppm)
Sulfur Dioxide (SO2)	0.13 ppm	Parts per million (ppm)
Temperature (TMP)		Celsius Grades (C)
Relative Humidity (RH)		Percent (%)
Wind speed (WSP)		Meters over second (m/s)
Wind Direction (WDR)		North Grados

Table 1. Pollutants considered in this study.

Initially, we obtain a database with 140,000 instances. The first step in our study consists of cleaning the initial dataset, with the aim to remove records with missing values, inconsistent values, and non-relevant information. This preprocessing step is explained in the next section.

4 Preprocessing Step

4.1 Data Cleaning and Missing values and Data Integration

The original database recollected in our work included missing values, non-relevant information (other chemical variables), data from other stations (not considered in this work), changes in standards of measurement, missing values, and noise. Moreover, at the beginning the original data was not labeled.

We started filtering the data, using a program implemented in R software [12]. In this software, it is possible to remove irrelevant data, including dependencies between records and values that were not in a standard.

After that, each record was labeled with a class. This process was performed considering the level of ozone reported in the NOM-1993[13]. Table 2 shown the different range of values admitted by the Mexican norm.

O_3	$Class$	Air quality
0.000-0.055	Green	Good
0.056-0.110	Yellow	Regular
0.111-0.165	Orange	Bad
0.166-0.220	Red	Very Bad
> 0.220	Purple	highly bad

Table 2. Classification of ozone ranges based on NOM-1993.

Although it is known that the used classifiers work well with missing values, an estimation was performed to approach that values. Consider that we need to estimate the missing value V_n for the record n. then that value is calculated with the average between its corresponding neighbors, as is shown in Ec. 1:

$$V_n = \frac{V_{n-1} + V_{n+1}}{k} \tag{1}$$

where $k = 2$. If values V_{n-1} or V_{n+1} are missing too, then are considered the next neighbors with known values. This process is possible to perform because each record was obtained hour by hour.

4.2 Descriptive statistics

The analysis based on descriptive statistics is important because allow us to identify general behavior of our data. We use this type of analysis with the aim to know trends of air pollutants between 2010 and 2011. In Table 5 is presented a summary of our dataset, considering each pollutant.

We can see that Pedregal was the highest contaminated area with respect to the others two. The total number of dangerous events (over two years) where acceptable levels of ozone were overcome (DEO) was 189 (according with the

	Pedregal (DSN=189)			Tlalnepantla (DSN=121)			Xalostoc (DSN=86)		
Cont.	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.
O_3	0.000	0.033	0.182	0.000	0.026	0.183	0.000	0.024	0.150
CO	0.000	0.515	2.900	0.000	0.920	5.100	0.000	1.102	11.30
NO_2	0.001	0.025	0.115	0.004	0.033	0.161	0.000	0.034	0.138
SO_2	0.000	0.005	0.097	0.000	0.009	0.283	0.000	0.007	0.143
TMP	1.00	16.08	31.50	1.60	17.92	36.00	1.200	17.50	33.80
RH	0.000	45.83	95.00	1.00	45.46	100.00	1.00	46.80	100.00
WDR	0.000	-	360	0	-	360	0	-	360
WSP	0.000	1.80	7.60	0.00	2.00	8.20	0.00	2.0	10.90

Table 3. Descriptive statistics summary for each meteorological station.

Mexican Norm). Based on this statistics, we can see that the mean of the ozone was 0.033 ppm, with a variation between 0 ppm to 0.180 ppm.

Tlalnepantla was the second place with high ozone levels. Between 2010 and 2011, the total number of times that the NOM was exceeded is 189, with a mean of 0.033 ppm and a variation between 0 ppm to 0.180 ppm. Finally, Xalostoc (located at the north of Mexico city) had the lowest number of environmental contingencies, 86, with a mean of 0.024 ppm, and variations between 1.6 ppm to 0.183 ppm.

Initially, our data include 8 variables or attributes, which are considered predictors of Ozone according to the literature. However is important to determine that this theory is correct. In the next section is presented an attribute selection process, with the aim to detect the most important attributes.

4.3 Attribute Selection

Before to proceed with the training step, it is necessary to know the most important attributes that could influence the levels of ozone, and simultaneously independent between them. For this task, we decide to work with the attribute evaluation based on the Chi-Square measure. This process was computed using the implementation available in Weka [3]. After to process our dataset, the results are shown in Table 4.

Attribute	PEDREGAL	TLALNEPANTLA	XALOSTOC
HOUR	3.8+-0.4	3.2+-0.7	3.3+-0.4
CO	3.1+-0.5	2.9+-0.9	2.0+-0.0
NO_2	5.9+-0.3	3.7+-1.3	6.1+-0.8
SO_2	2.1+-0.3	4.8+-1.6	6.5+-2.2
TMP	1.0+-0.0	1.0+-0.0	1.0+-0.0
RH	5.1+-0.3	5.6+-0.4	5.0.+-0.6
WDR	7.6+-0.4	7.9+-0.3	7.3+-0.4
WSP	7.4+-0.4	6.9+-0.3	4.8+-0.7

Table 4. Ranking obtained with the χ^2 test with Weka.

These results can be interpreted as follows: the most significant attributes are ranked with a value closer to 1, therefore the worst attribute is ranked with a value farthest to 1. It is easy to see that the best attribute for Pedregal, Tlalnepantla and Xalostoc is TMP, followed by the chemical pollutants: SO_2, NO_2 and CO, and finally the attributes RH, WSP and WDR. Considering these results, we can select the best attributes for our experiments.

4.4 Building the training set

Since we need a training set, we perform a stratified sampling over our data, where the strata is based on the seasons: spring, summer, fall, and winter. This process was implemented with a script based on "AWK". In general, our routine select (or not select) a tuple in our dataset based on a value computed with a random function. If that value exceeds a threshold (defined previously by the user), then the tuple is selected and assigned a value: spring, summer, fall or winter, according with its date.

As result of this process, only 30% of the tuples in the original dataset were selected, with 14000 tuples per each meteorological station. However, based on a statistical analysis, the selected dataset is unbalanced. Then we perform a process with the aim to balance the selected dataset, decreasing the total number of tuples with the majority class, and increasing the records with the minority class.

After that, we have a dataset ready to be analyzed with the machine learning algorithms, such as neural networks and decision trees. Our experiments are exposed in the next section.

5 Experimental Results

For the training phase, we performed a set of experiments with multilayer perceptron neutral network and decision trees, all with teen fold cross validation. For the case of neural networks, several experiments were conducted with different topologies (including from 3 to 9 neurons per layer, with 1, 2, and 3 hidden layers), using a learning rate = 0.3, and momentum = 0.2 with a sigmoidal function.

With respect to the neural networks, our results are shown in Table 5. In this table, the results are divided in 500, 1000 and 2000 epochs.

We can see that the best configuration for the Pedregal is obtained using one hidden layer with eight neurons, and 500 epochs. For this case, the accuracy is 87.7%. On the other hand, in the case of Tlalnepantla the best configuration is neural network with one hidden layer, and nine neurons, with 500 epochs. Finally, for Xalostoc the highest accuracy was 85.3%.

For the case of decision trees, we use five different algorithms: C4.5, Random Tree, BF Trees, Decision Stump and Random Forest. The results obtaining in our experiments are show in Table 6. We proposed different configurations for each meteorological site. First, with the C4.5 algorithm, we use the following

		1 HL.			2 HL.			3 HL.		
#Neu.	**Epochs.**	Ped.	Tla.	Xal.	Ped.	Tla.	Xal.	Ped.	Tla.	Xal.
3	500	84.9	81.7	89.6	83.3	82.0	87.6	83.5	81.3	88.2
	1000	85.1	81.9	89.4	84.5	81.6	87.7	83.6	80.4	88.3
	2000	85.3	81.1	89.3	83.4	81.2	87.7	84.5	80.0	88.1
4	500	87.3	85.0	91.9	83.9	84.3	90.1	84.8	83.7	90.5
	1000	86.7	85.2	91.8	83.6	84.1	90.3	84.1	83.3	90.9
	2000	87.1	85.1	92.0	84.3	84.4	91.8	84.7	83.5	90.7
5	500	87.3	85.2	92.1	84.9	85.6	90.7	83.8	84.6	90.9
	1000	86.6	85.1	92.0	84.9	84.6	91.2	84.5	84.4	90.9
	2000	86.4	85.0	91.9	84.8	84.8	91.0	84.7	85.1	90.7
8	500	88.7	85.6	93.2	86.5	85.1	92.2	85.3	84.2	91.4
	1000	87.7	85.1	93.2	87.3	85.7	91.8	85.0	85.0	90.7
	2000	85.0	85.5	93.2	86.0	85.4	91.8	85.4	83.9	91.4
9	500	87.6	86.7	93.6	86.5	85.3	91.9	84.2	85.2	91.7
	1000	88.6	86.4	93.5	86.7	84.3	92.1	84.5	84.7	92.1
	2000	88.0	86.1	93.3	85.9	85.4	92.2	84.9	86.0	91.9
Avg.	500	86.4	85.0	92.6	86.2	85.1	91.5	82.9	84.5	91.1

Table 5. Results of Multilayer perceptron using different configurations.

configuration: $factor = 0.05$, $MinNumObj = 2$ and $Unpruned = FALSE$. For the case of Random Trees we used the following parameters: $K = 0$, $numFolds = 0$ and $seed = 1$. In the case of BFTree, the parameters was adjusted as follows: $minNumObj = 2$, $numFoldsPrunning = 5$, $seed = 1$ and $sizePer = 1$. We use the default configuration for the case of Decision Stump tree. Furthermore for Random Forest, the parameters were: $MaxDepth = 10$, $Debug = False$, $NumTrees = 50$, $Seed = 1$. In Table 6 are shown our results.

Algorithm	Ped.	Tla.	Xal
C4.5	91.6	88.2	93.3
Random Forest	92.3	89.6	94.4
Random Tree	88.3	85.4	92.0
BF Tree	89.0	85.2	92.6
Decision Stump	70.0	72.0	86.30

Table 6. Results of decision trees.

In our results, we can see that the best accuracy for the three meteorological stations Pedregal, Tlalnepantla and Xalostoc is obtained with Random Forest: 92.3%, 89.6% and 94.4% respectability, but the precision of C4.5 is very similar.

In general, we can see that the accuracy between neural networks (multilayer perceptron) and decision trees (C4.5 and Random Forest) in very closer, with small differences. However, decision trees are descriptive models. For some cases, it is very helpful to describe the patterns, and can be used to construct decision rules.

Based on the above mentioned, the accuracy obtained in our experiments are very promising for predicting high levels of ozone.

6 Conclusions and future work

The machine learning techniques used for forecast purposes can be a very useful tool in decision making, planning and evaluation of air quality. In this paper we present two models for air quality forecast: the first based on decision tree and the second based on neural networks. In our experiments we used data comes from of the SIMAT meteorological network. According with our results, it is possible to predict high levels of ozone. Finally, as future work we will analyze effects that have other chemical variables in the levels of ozone, such as: particles less than 10 micrograms (NOx), pluvial precipitation (P_v), solar radiation, and so on.

References

1. Reyes, H., Vaquera, H., or, J.V.: Estimate of tendencies in high levels of urban ozone using the quantiles of the distribution of generalized extreme values (gev). (en revisin) (2007)
2. SIMAT, D.: Biblioteca Virtual. www.sma.df.gob.mx/simat, México (2012)
3. Mark Hall, Eibe Frank, G.H.B.P.P.R.I.H.W.: The weka data mining software: An update; sigkdd explorations, volume 11, issue 1. Master's thesis, Tesis de Maestría en Ingeniera Ambiental-UNAM (2012)
4. Bravo, H.: La contaminación atmosférica por ozono en la zona metropolitana de la ciudad de méxico: evolución histórica y perspectivas. IX Comisión Nacional de los Derechos Humanos (1992)
5. I.N.E.G.I.: Estadísticas del Medio Ambiente. Semarnap, Mxico (1999)
6. Molina, M., L., M.: The impacts of magacities on air pollution, environmetal aspects of urbanization, Goteborg Sweden (2004)
7. Garfias, M., Audry, J., Garfias, F.: Ozone trend analysis at pedregal station in the metropolitan area of mexico city. J. Mex. Chem. Soc. **49(4)** (2005) 322–323
8. Seinfeld, J.: Comittee on tropospheric ozone formation and measurement; Board on Environmente Studies and Toxicology; Board on Atmospheric Sciences and Climate; Commission on Geosciencies, Environment and Resources; National Research Council, Rethinking the on ozone problem in urban and regional air Pollution. National Academic Press, Washington (1991)
9. E., Aguirre, A.A.L.B.: A sistem for forecast of the maximun ozone levels. Atmospheric Environment **38** (2004) 4689–4699
10. Barai, S., Dikshit, A., Sharma, S.: Neural network models for air quality prediction: A comparative study. In Saad, A., Dahal, K., Sarfraz, M., Roy, R., eds.: Soft Computing in Industrial Applications. Volume 39 of Advances in Soft Computing. Springer Berlin / Heidelberg (2007) 290–305
11. Mitchell, T.M.: Machine Learning. 1 edn. McGraw-Hill, Inc., New York, NY, USA (1997)
12. Proyect, R.: Biblioteca Virtual. http://www.r-project.org/, USA (2012)
13. D.F., G. Gaceta Oficial No. 129 **129** (2006) 1–14