27

Forecasting Air Quality Data with the Gamma Classifier

Itzamá López-Yáñez[1], Cornelio Yáñez-Márquez[1],
Víctor Manuel Silva-García[2]
[1] *IPN Computing Research Center,*
[2] *IPN Computing Innovation and Technological Development Center*
Mexico

1. Introduction

Environmental topics have gained the attention of increasingly large portions of global population. In different languages and through diverse means, civil associations launch campaigns for people to realize the importance of protecting the environment (Toepfer *et al*., 2004; Hisas *et al*., 2005), even attracting the active participation of governments (United Nations, 1992; United Nations, 1997; Secretaría de Comercio y Fomento Industrial, 1986; Web del Departamento de Medio Ambiente y Vivienda de la Generalitat de Cataluña, 2007).
Computer Sciences have not been immune to the awareness dawn. In this sense, several techniques of artificial intelligence have been applied to the analysis and forecasting of environmental data, such as artificial neural networks (Sucar *et al*., 1997; Dutot *et al*., 2007; Salazar-Ruiz *et al*., 2008) and Support Vector Machines (Wang *et al*., 2008). One particular technique which has been recently used in the prediction of environmental data —in particular, air quality data— is the Gamma classifier (López, 2007). This relatively new algorithm has shown some promising results.
In this work the Gamma classifier is applied to forecast air quality data present in public databases measured by the Mexico City Atmospheric Monitoring System (*Sistema de Monitoreo Atmosférico*, SIMAT in Spanish) (Sistema de Monitoreo Atmosférico de la Ciudad de México, 2007).
The rest of the chapter is organized as follows: the air quality data and SIMAT are described in section 2, while section 3 is dedicated to the Gamma classifier. Section 4 contains the main proposal of this work, and in section 5 the experimental results are discussed. Conclusions and future work are shown in section 6.

2. SIMAT

The Mexico City Atmospheric Monitoring System (*Sistema de Monitoreo Atmosférico de la Ciudad de México*, SIMAT in Spanish) is tightly coupled with the evolution of the Mexican

capital, and with the problems inherent to its development. The information herein presented is taken from (Sistema de Monitoreo Atmosférico de la Ciudad de México, 2007). SIMAT is committed to operate and maintain a trustworthy system for the monitoring of air quality in Mexico City, as well as analyzing and publishing this information in order to fulfil the current requirements and legislation. The objective of SIMAT is to watch and evaluate the air quality in Mexico City, as a pre-emptive measure for health protection of its inhabitants, in order to promptly inform the populace as well as enable decision making in prevention and air quality improvement programs. SIMAT is made up by four specialized subsystems, one Atmospheric Monitoring Mobile Unit, and a Calibration Standards Transfer Laboratory. The four subsystems are:

- RAMA (Automatic Atmospheric Monitoring Network, *Red Automática de Monitoreo Atmosférico* in Spanish) takes continuous and permanent measurements of several contaminants: ozone (O_3), sulphur dioxide (SO_2), nitrous oxides (NO_x), carbon monoxide (CO), particulate matter less than 10 microns in diameter (PM_{10}), and particulate matter less than 2.5 microns in diameter ($PM_{2.5}$); each measurement is taken automatically every hour.
- REDMA (Manual Atmospheric Monitoring Network, *Red Manual de Monitoreo Atmosférico* in Spanish) monitors particulate matter suspended in the air — particulate matter less than 10 microns in diameter (PM_{10}), particulate matter less than 2.5 microns in diameter ($PM_{2.5}$), and total suspended particulate matter (PST) —, as well as their concentration and composition; each measurement is taken manually every six days.
- REDMET (Meteorological and Solar Radiation Network, *Red de Meteorología y Radiación Solar* in Spanish) monitors meteorological parameters —such as wind direction and speed— and solar radiation, in order to elaborate meteorological forecasting and dispersion models; it also records and monitors the UV index.
- REDDA (Atmospheric Deposit Network, *Red de Depósito Atmosférico* in Spanish) measures both dry and wet deposit, whose analysis allows the study of rain properties and the flow of toxic substances from the atmosphere to the surface.

IMECA	Condition	Effects on Health
0-50: green	Good	Suitable for conducting outdoor activities
51-100: yellow	Regular	Possible discomfort in children, the elderly and people with illnesses
101-150: orange	Bad	Cause of adverse health effects on the population, particularly on children and older adults with cardiovascular and / or respiratory illnesses such as asthma
151-200: red	Very Bad	Cause of greater adverse health effects on the population, particularly on children and older adults with cardiovascular and / or respiratory illnesses such as asthma
>200: purple	Extremely Bad	Cause of adverse health effects in the general population. Serious complications may present in children and older adults with cardiovascular and / or respiratory illnesses such as asthma

Table 1. IMECA and its implications for health

The Air Quality Metropolitan Index (*Índice Metropolitano de la Calidad del Aire*, IMECA in Spanish) is a reference value for people to be aware of the pollution levels prevalent in any zone, in a precise and timely manner, in order to take appropriate protection measures. When the IMECA of any pollutant is greater than 100 points, its concentration is dangerous for health and, as the value of IMECA grows, the symptoms worsen, as can be seen in table 1.

Generating the IMECA is one of the primordial tasks of SIMAT. Since July 1st, 1998, the IMECA has been transmitted 24 hours every day to different electronic and printed communication media. Currently, the hourly value of IMECA can be consulted online in (Sistema de Monitoreo Atmosférico de la Ciudad de México, 2007) and also by telephone at the IMECATEL service, which started operations on March 22, 2001. On both of these services, information is available 24 hours a day.

In November 2006, the *Gaceta Oficial del Distrito Federal* published the Federal District Environmental Norm (*Norma Ambiental para el Distrito Federal*) NADF-009-AIRE-2006 (Gobierno del Distrito Federal, 2006), which states the specifications for elaborating the IMECA for the criteria pollutants, such as: O_3, NO_2, SO_2, CO, PM_{10} and $PM_{2.5}$.

For each of the criteria pollutants, the norm states equations for calculating the corresponding IMECA, from the concentration data. Tables 2, 3, and 4 show these equations for CO, O_3, and SO_2, respectively. With these equations, the IMECA value and IMECA level (condition) can be easily computed from the concentration of each of the pollutants, in parts per million (ppm).

IMECA Interval	Concentration Intervals (ppm)	Equations
0-50: green	0-5.50	$\mathrm{IMECA}[\mathrm{CO}] = \mathrm{con}[\mathrm{CO}] \times 50/5.50$
51-100: yellow	5.51-11.00	$\mathrm{IMECA}[\mathrm{CO}] = 1.82 + \mathrm{con}[\mathrm{CO}] \times 49/5.49$
101--150: orange	11.01-16.50	$\mathrm{IMECA}[\mathrm{CO}] = 2.73 + \mathrm{con}[\mathrm{CO}] \times 49/5.49$
151--200: red	16.51-22.00	$\mathrm{IMECA}[\mathrm{CO}] = 3.64 + \mathrm{con}[\mathrm{CO}] \times 49/5.49$
>200: purple	>22.00	$\mathrm{IMECA}[\mathrm{CO}] = \mathrm{con}[\mathrm{CO}] \times 201/22.01$

Table 2. IMECA calculation equations for carbon monoxide (CO)

IMECA Interval	Concentration Intervals (ppm)	Equations
0-50: green	0-0.055	$\mathrm{IMECA}[\mathrm{O_3}] = \dfrac{\mathrm{con}[\mathrm{O_3}] \times 100}{0.11}$
51-100: yellow	0.056-0.110	
101--150: orange	0.111-0.165	
151--200: red	0.166-0.220	
>200: purple	>0.220	

Table 3. IMECA calculation equation for ozone (O_3)

IMECA Interval	Concentration Intervals (ppm)	Equations
0-50: green	0-0.065	$\mathrm{IMECA}[SO_2] = \frac{\mathrm{con}[SO_2] \times 100}{0.13}$
51-100: yellow	0.066-0.130	
101--150: orange	0.131-0.195	
151--200: red	0.196-0.260	
>200: purple	>0.260	

Table 4. IMECA calculation equation for sulphur dioxide (SO_2)

3. The Gamma classifier

This pattern classifier, of recent proposal, has shown some very promising results. The following discussion is strongly based on (López, 2007).

The basis of the Gamma classifier is the gamma operator, hence its name. In turn, the gamma operator is based on the a, β, and u_β operators and their properties, in particular when dealing with binary patterns coded with the modified Johnson-Möbius code. Also, it is important to define the sets A and B, since they are used throughout this work. Thus, let there be the sets A = {0, 1} and B = {0, 1, 2}.

3.1 Preliminaries

The alpha and beta operators are defined in tabular form, taking into account the definitions of the sets A and B, as shown in table 5.

$\alpha : A \times A \to B$		
x	y	$\alpha(x, y)$
0	0	1
0	1	0
1	0	2
1	1	1

$\beta : B \times A \to A$		
x	y	$\beta(x, y)$
0	0	0
0	1	0
1	0	0
1	1	1
2	0	1
2	1	1

Table 5. Definition of the Alpha and Beta operators

Now, the unary operator u_β, which receives as input an n dimensional binary vector **x**, and outputs a non-negative integer number, is calculated as shown in equation 1.

$$u_\beta(\mathbf{x}) = \sum_{i=1}^{n} \beta(x_i, x_i) \tag{1}$$

On the other hand, the modified Johnson-Möbius code allows to convert a set of real numbers into binary representations by following these steps:

1. Substract the minimum (of the set of numbers) from each number, leaving only non-negative real numbers.

2. Scale up the numbers (truncating the remaining decimals if necessary) by multiplying all numbers by an appropriate power of 10, in order to leave only non-negative integer numbers.
3. Concatenate e_m–e_j zeros with e_j ones, where e_m is the greatest non-negative integer number to be coded, and e_j is the current non-negative integer number to be coded.

Finally, the generalized gamma operator γ_g, which takes as input two binary patterns $\mathbf{x} \in A^n$ and $\mathbf{y} \in A^m$, with $n, m \in \mathrm{Z}^+$, $n \leq m$, and a non-negative integer number θ; and gives a binary number as output; can be calculated as in equation 2.

$$\gamma_g(\mathbf{x}, \mathbf{y}, \theta) = \begin{cases} 1 & \text{if } m - u_\beta[\alpha(\mathbf{x}, \mathbf{y}) \bmod 2] \leq \theta \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

With these tools we are now ready to present the algorithm for the Gamma classifier.

3.2 The Gamma classifier algorithm

Let $k, m, n, p \in Z^+$; $\{\mathbf{x}^\mu | \mu = 1, 2, \ldots, p\}$ be the fundamental pattern set with cardinality p, where $\forall \mu \, \mathbf{x}^\mu \in \mathrm{R}^n$, and let $\mathbf{y} \in \mathrm{R}^n$ be an n dimensional real-valued pattern to be classified. It is assumed that the fundamental set is partitioned into m different classes, each class having a cardinality k_i, $i = 1, 2, \ldots, m$, thus $\sum_{i=1}^{m} k_i = p$. In order to classify $\mathbf{y}$, these steps are followed:

1. Code the fundamental set with the modified Johnson-Möbius code, obtaining a value e_m for each component. The e_m value is calculated as defined in equation 3.

$$e_m = \bigvee_{i=1}^{p} x_j^i \tag{3}$$

2. Compute the stop parameter, as expressed in equation 4.

$$\rho = \bigwedge_{j=1}^{n} e_m(j) \tag{4}$$

3. Code $\mathbf{y}$ with the modified Johnson-Möbius code, using the same parameters used with the fundamental set. If any y_j is greater than the corresponding $e_m(j)$, the γ_g operator will use such y_j instead of m.
4. Transform the index of all fundamental patterns into two indices, one for the class they belong to, and another for their position in the class (i.e. $\mathbf{x}^\mu$ which belongs to class i becomes $\mathbf{x}^{i\omega}$).
5. Initialize θ to 0.

6. Do $\gamma_g\left(\mathbf{x}_j^{i\omega}, \mathbf{y}_j, \theta\right)$ for each component of the fundamental patterns in each class, following equation 2.
7. Compute a weighted sum c_i for each class, according to equation 5.

$$c_i = \frac{\sum_{\omega=1}^{k_i}\sum_{j=1}^{n}\gamma_g\left(\mathbf{x}_j^{i\omega}, \mathbf{y}_j, \theta\right)}{k_i} \tag{5}$$

8. If there is more than one maximum among the different c_i, increment θ by 1 and repeat steps 6 and 7 until there is a unique maximum, or the stop condition $\theta \geq \rho$ is fulfilled.
9. If there is a unique maximum, assign $\mathbf{y}$ to the class corresponding to that maximum:

$$C_y = C_j \quad \text{such that} \quad \bigvee_{i=1}^{m} c_i = c_j \tag{6}$$

10. Otherwise, assign $\mathbf{y}$ to the class of the first maximum.

The Gamma classifier is inspired on the Alpha-Beta associative memories, taking the alpha and beta operators as basis for the gamma operator. As such, the Gamma classifier is a member of the Associative Approach to Pattern Recognition, in which the algorithms and models use concepts and techniques derived from associative memories in order to recognize and classify patterns.

As can be seen, this classifier is relatively simple, requiring simple operations. Its complexity is polynomial, as was shown in (López, 2007). Also, notice that while being iterative, the classifier will reach a solution in finite time: at best in one iteration, at worst in the same amount of iterations as the stop parameter indicates (see equation 4).

Although the gamma classifier is not old, it has already been applied to several different problems: classification of the Iris Plant database, localization of mobile stations, software development effort estimation of small programs, and of course environmental data prediction. In these problems, some quite different from each other —and even unfulfilling of the basic premises of the classifier—, the Gamma classifier has shown competitive experimental result.

4. Proposed application

In the current work, the authors apply the Gamma classifier to environmental data obtained from databases derived from SIMAT, in order to forecast air quality data. In particular, the RAMA database was used. The experiments were conducted on three pollutants: carbon monoxide (CO), ozone (O_3), and sulphur dioxide (SO_2). The fundamental set was built with all the samples taken at a particular monitoring station during 2006, while the testing set was built with the samples taken during two non-consecutive months of 2007: February and May. Given that not all stations sample all pollutants, different stations were selected for each pollutant: IMP (*Instituto Mexicano del Petróleo*) for CO, CES (*Cerro de la Estrella*) for O_3, and TLI (*Tultitlán*) for SO_2. This can be seen also in table 6.

		Fundamental set			Testing set		
Experiment	**Pollutant**	**Period**	**Station**	**Size**	**Period**	**Station**	**Size**
1	CO	2006	IMP	8710	2007-Feb	IMP	651
2	CO	2006	IMP	8710	2007-May	IMP	723
3	O3	2006	CES	8749	2007-Feb	CES	651
4	O3	2006	CES	8749	2007-May	CES	723
5	SO2	2006	TLI	8749	2007-Feb	TLI	641
6	SO2	2006	TLI	8749	2007-May	TLI	711

Table 6. Composition and sizes of fundamental and testing sets for each experiment

Each pattern is made up by n successive samples, concatenated each after the other. As the class for such pattern, the n+1–th sample is used. Thus, patterns are built from the samples as mentioned above, and then these patterns are grouped together in fundamental and testing set for each experiment. The composition and size of each of these set, for each experiment, can be seen in table 6.

5. Experimental Results

As mentioned in the previous section, both the fundamental set and the testing set were formed with data taken from the RAMA database for each pollutant, containing hourly samples of concentration measured in parts per million (ppm).
With these data, input patterns of 10 samples were formed; that is, n=10. While the value of n can be arbitrarily chosen, 10 gave good results in preliminary tests. The output patterns (i.e. the class) were taken from the sample following the last sample in the pattern.
Once trained with the fundamental set, the Gamma classifier is presented with the testing set, obtaining the pollutant concentration forecast for the next hour (see figures 1 through 6).

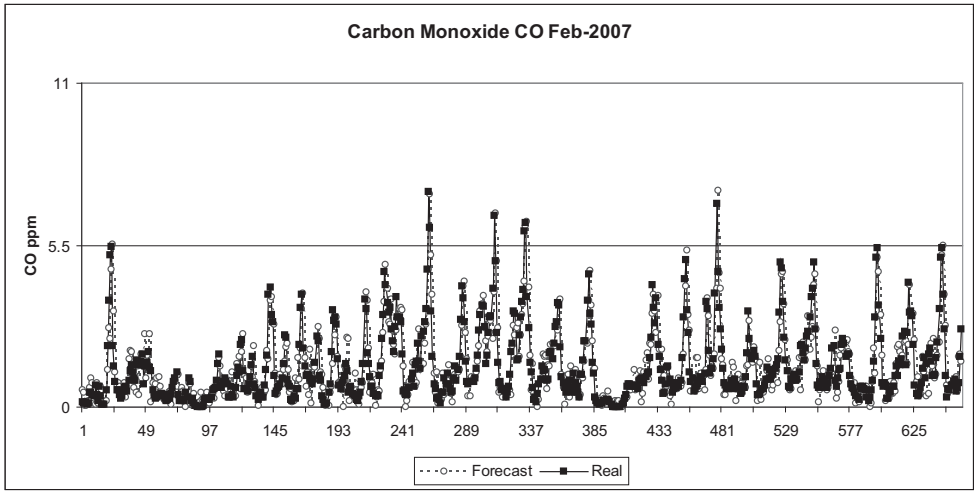

Fig. 1. Predicted values *vs* real values for carbon monoxide (CO), February 2007

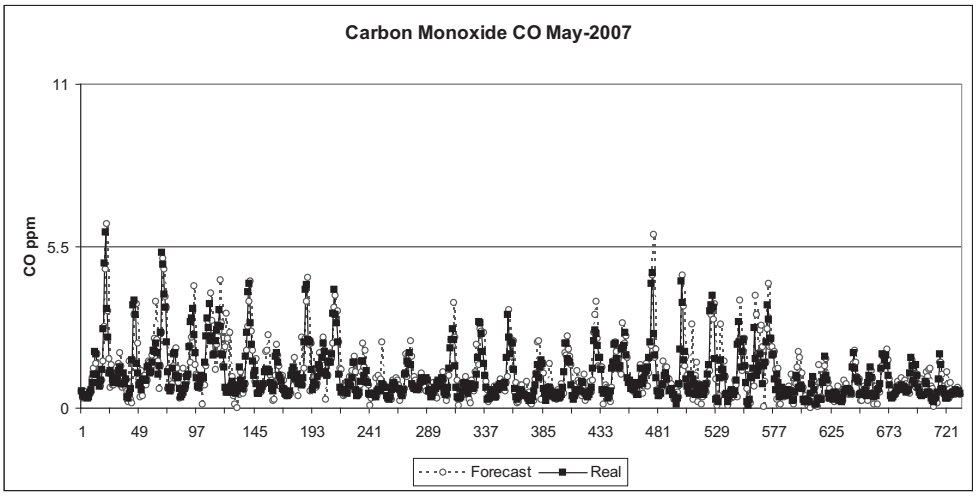

Fig. 2. Predicted values *vs* real values for carbon monoxide (CO), May 2007

Here are some examples of the results obtained: for experiment 1 (CO Feb 2007) on February 3rd at 18:00, the measured (real) CO concentration was 0.42 ppm, while the Gamma classifier predicted 0.42 ppm, which gives an error of 0.00 ppm. While this is clearly the best result, some error can be found too. For experiment 4 (O_3 May 2007) on May 12 at 17:00 the system predicted 0.034 ppm of O_3 concentration, while the observed value was 0.048 ppm, for an error of -0.014 ppm. Yet larger errors can be seen; for instance, on experiment 5 (SO_2 Feb 2007) February 19 at 1:00, the forecast was 0.059 ppm while the real concentration of SO_2 was 0.251, which amounts to an error of -0.192 ppm. These examples can be seen in table 7.

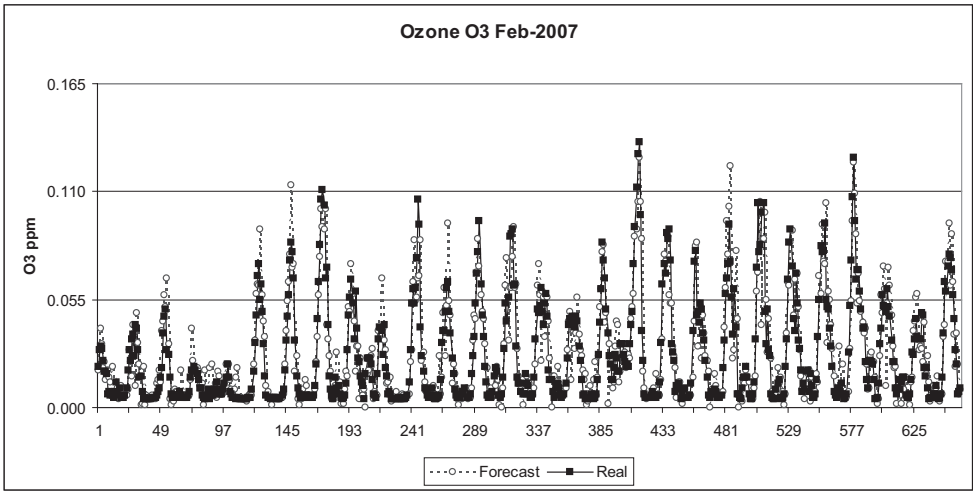

Fig. 3. Predicted values *vs* real values for ozone (O_3), February 2007

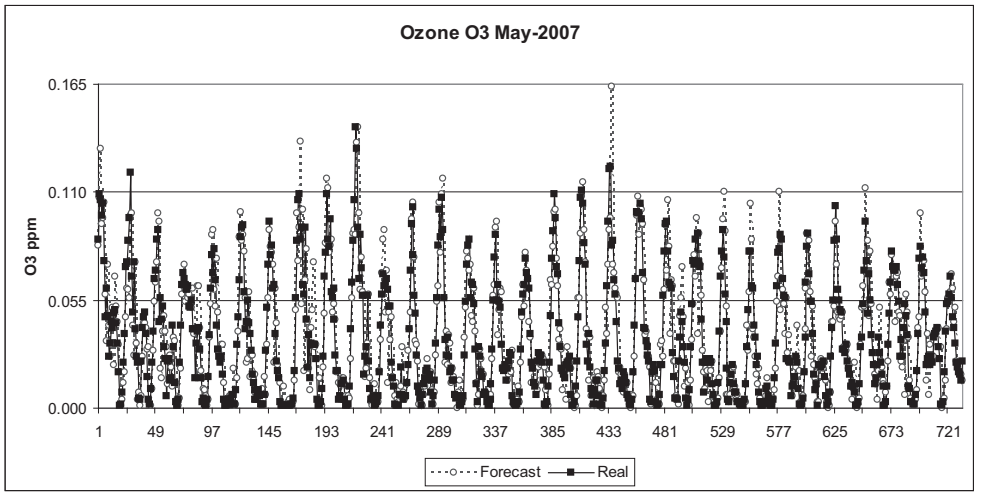

Fig. 4. Predicted values *vs* real values for ozone (O_3), May 2007

Pollutant	Date	Hour	Forecast	Observation	Error
CO	February 3	18:00	0.42 ppm	0.42 ppm	0.00 ppm
O_3	May 12	17:00	0.034 ppm	0.048 ppm	-0.014 ppm
SO_2	February 19	1:00	0.059 ppm	0.251 ppm	-0.192 ppm

Table 7. Examples of results

An interesting characteristic of these pollutants, which can be observed in the figures (1 through 6), is that CO and O_3 have a periodic behaviour according to the hour of the day,

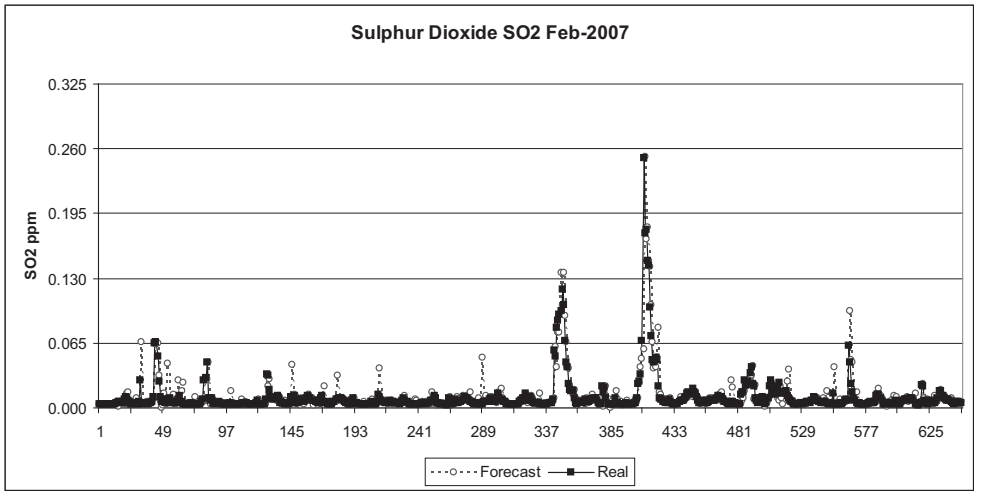

Fig. 5. Predicted values *vs* real values for sulphur dioxide (SO_2), February 2007

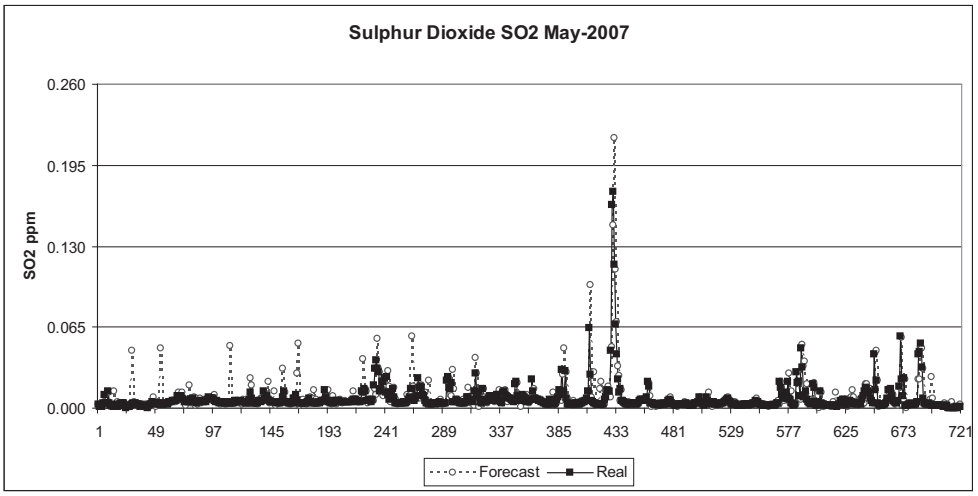

Fig. 6. Predicted values *vs* real values for sulphur dioxide (SO_2), May 2007

while SO_2 does not present such an easily discernible periodic behaviour. In the case of CO, the peaks usually happen in the second quarter of the day (6:00 to 12:00 hours), while in the case of O_3, the peaks occur more commonly around the third quarter of the day (12:00 to 18:00 hours). It is also noteworthy that the greater errors usually appear close to a peak, either positive or negative; this last observation is true for all three pollutants.

An example of the latter observation is that on February 15 and 18, there was a sharp change in the behaviour of SO_2: while during the rest of the month the concentration of this pollutant was low (it remained in the 0-0.65 ppm range, indicating a good IMECA condition), in those days the concentration reached 0.119 ppm (Feb. 15) and 0.251 (Feb. 18) for an IMECA condition of very bad. Again, on May 19, the SO_2 concentration reached exceptional levels: 0.174 ppm when the mean for that month was 0.007 ppm. It is clear that these were exceptional situations, which lie out of the ordanary situations. If the Gamma classifier (or any other algorithm for that matter) is not trained with such exceptional ciscumstances, it is to be expected that the corresponding forecast will not be accurate.

Two quantitative measures of the performances shown by the Gamma classifier on this application were used. On one side, Rooted Mean Square Error (RMSE), which is a widely used measure of performance and is calculated as shown in equation 7. On the other side, the bias, which can be calculated by following equation 8, is used to describe how much the system is underestimating or over estimating the results. For both equations, P_i is the *i*–th predicted value and O_i is the *i*–th observed (real) value.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_i - O_i\right)^2} \tag{7}$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}\left(P_i - O_i\right) \tag{8}$$

The RMSE and bias exhibited by the results of each experiment are shown in table 8. Notice that the bias is small in all cases, especially if the size of the testing sets (641 patterns in the smallest, 723 in the largest) is taken into account.

Experiment	Pollutant	Station	Testing Period	RMSE	Bias
1	CO	IMP	2007-Feb	0.726013	7.96
2	CO	IMP	2007-May	0.611769	45.58
3	O_3	CES	2007-Feb	0.012302	0.607
4	O_3	CES	2007-May	0.014443	0.306
5	SO_2	TLI	2007-Feb	0.012096	0.573
6	SO_2	TLI	2007-May	0.010487	0.439

Table 8. RMSE and bias for each experiment

The RMSE exhibited by the three pollutants is also small. Again, that of CO is comparatively larger, although when the order of the data processed is taken into account (the mean of CO concentration during 2006 was 1.237 ppm), having a RMSE of 0.726 ppm for February and one of 0.612 ppm for May is relatively small. Ozone exhibited smaller values for RMSE: 0.0123 ppm for February and 0.0144 ppm for May. If these values are compared to the average O_3 concentration for 2006, 0.0262 ppm, it is clear that the error shown is not too high. The smallest RMSE of the three pollutants was presented by SO_2: 0.0121 ppm on February and 0.0105 ppm on May. The annual mean for this pollutant in 2006 was 0.0099 ppm, which is smaller than the RMSE for both experiments.

These comparisons between the RMSE and the previous annual mean for each pollutant is not a particularly good measure of how good the prediction was. They only indicate that the errors are close to said mean, preferably smaller. A better measure would be to compare these results with those offered by using other methods.

However, there is a problem with such comparison too. Most authors use data taken from databases close to them. Therefore, most results are obtained by processing data different for each method; it even happens that different databases use different units to measure the same pollutants [!]. Thus, a direct comparison is not appropriate, even though the same measure of error is used. It is due to this lack of a standard, benchmarking database that comparisons should be done carefully.

One example is the comparison between the results presented in (Sucar *et al.*, 1997) and those shown in the current work. The experiments on both publications were done with the same database (SIMAT – RAMA), with the same pollutant (in the case of experiments 3 and 4: O_3), with the same unit of sample measure (ppm), and with the same measure of error (absolute average error). However, the data used for experimentation was taken from different stations and different years. Although the results are comparable, they are not directly comparable. And this is the best match of data; greater caution should be taken when comparing the results obtained by experiments done on data taken from different databases.

Table 9 presents a comparison between the results obtained in this work and those presented in other publications, with the restriction of using the same database: SIMAT; in particular, the RAMA database. Notice that the results presented in (Sucar *et al.*, 1997) are greatly surpassed.

Experiment	Algorithm Used	Pollutants Considered	Size of Training / Testing Sets	Performance (Abs. Avg. Error)
	Bayesian network (Sucar *et al.*, 1997)	O_3 (ppm)	400 / 200	0.221000
	Neural network (Sucar *et al.*, 1997)	O_3 (ppm)	400 / 200	0.160000
	C4.5 (Sucar *et al.*, 1997)	O_3 (ppm)	400 / 200	0.176400
	Gamma classifier (Yáñez-Márquez *et al.*, 2008)	SO_2 (ppm)	8749 / 709	0.000408
1	Gamma classifier (current work)	CO (ppm)	8710 / 651	0.012042
2	Gamma classifier (current work)	CO (ppm)	8710 / 723	0.062183
3	Gamma classifier (current work)	O_3 (ppm)	8749 / 651	0.000918
4	Gamma classifier (current work)	O_3 (ppm)	8749 / 723	0.000417
5	Gamma classifier (current work)	SO_2 (ppm)	8749 / 641	0.000676
6	Gamma classifier (current work)	SO_2 (ppm)	8749 / 711	0.000795

Table 9. Comparison of related results (SIMAT database) in absolute average error given for pollutant concentration

Experiment	Algorithm Used	Pollutants Considered	Size of Training / Testing Sets	Performance (RMSE)
	Neural network (Dutot *et al.*, 2007)	O_3 ($\mu g/m^3$)	613 / 105	15
	Neural network (Salazar-Ruiz *et al.*, 2008)	O_3 (ppb)	NA / 1343	9.43
			NA /2367	13.79
	Online SVM (Wang *et al.*, 2008)	SO_2 ($\mu g/m^3$)	240 / 168	12.96, 10.90
	CALINE3 (Gokhale & Raokhande, 2008)	PM_{10}, $PM_{2.5}$ ($\mu g/m^3$)	~120	88, 55
	Gamma classifier (Yáñez-Márquez *et al.*, 2008)	SO_2 (ppm)	8749 / 709	0.009218
2	Gamma classifier (current work)	CO (ppm)	8710 / 723	0.611769
3	Gamma classifier (current work)	O_3 (ppm)	8749 / 651	0.012302
6	Gamma classifier (current work)	SO_2 (ppm)	8749 / 711	0.010487

Table 10. Comparison of related results (diverse databases) in RMSE, given for pollutant concentration; NA indicates a not available value, ppb means parts per billion

Even taking into account the above mentioned discussion, the error exhibited by the Gamma classifier is several orders of magnitude smaller: the average error of experiment 3 (0.000918 ppm) is more than 150 times smaller than that of the neural network (0.160000 ppm); and experiment 3 did not give the best results [!]. Also, the results of experiments 3 and 4 (0.000417 ppm) are coherent with those of (Yáñez-Márquez *et al*., 2008) (0.000408 ppm), in the sense that they are quite similar.
On the other hand, table 10 shows the results of several experiments, done with data taken from different databases (one experiment was chosen for each pollutant among those presented in the current work). Taking these differences into consideration, as well as the fact that the results reported are based on different units (some in parts per million, other in parts per billion, and yet others in $\mu g/m^3$), it can be said that the Gamma classifier exhibits competitive performance.

6. Conclusions and Future Work

In this work, the utility of applying the Gamma classifier to forecasting air quality data has been experimentally shown. More specifically, the hourly concentration of three pollutants: carbon monoxide, ozone, and sulphur dioxide, as taken from the RAMA database, was analyzed. Six experiments were done, two on each pollutant: year 2006 was learned, and the hourly concentration values for the months of February 2007 and May 2007 were predicted.
The experimental results show a small error when compared to the data being predicted. However, it is noteworthy that most significant errors occur when the graph of the data changes direction (i.e. starts decreasing after increasing, or vice versa), implying a quite likely venue of improvement.
It is also clear that there were exceptional situations present in the data from which the testing sets for experiments 5 and 6 were built. In particular, the days February 15 and 18, and again in May 19, presented SO_2 concentration values which were out of the ordinary. Although these exceptional situations caused the Gamma classifier to incur on large errors for those days, the RMSE for both experiments was still competitive.
One possibility to improve the forecast results when such situations arise is to train the system with data taken during events similar to these, when the conditions were anomalous.
A different approach to improve the results, and not just for these experiments, would be to take into account several variables at the same time. For instance, learn from the data taken at several monitoring stations, or learning from several (related) pollutants at the same time.
It is also worthy of mention that direct comparisons with results reported in other works are difficult, since the same data is seldom used for experimentation on different works. It would greatly improve these comparisons to have a standard database to serve as a benchmark.

8. References

Dutot, A.-, Rynkiewicz, J., Steiner, F.E., Rude, J. (2007). A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions. *Environmental Modelling and Software*, Vol. 22, No. 9, (September 2007) 1261-1269, ISSN: 1364-8152.

Gobierno del Distrito Federal (2006). Norma Ambiental para el Distrito Federal (in Spanish). *Gaceta Oficial del Distrito Federal*, XVI Epoch.

Gokhale, S., Raokhande, N. (2008). Performance evaluation of air quality models for predicting PM10 and PM2.5 concentrations at urban traffic intersection during winter period. *Science of the Total Environment*, Vol. 394, No. 1 (May 2008) 9-24, ISSN: 0048-9697.

Hisas, Liliana *et al*. (2005). *A Guide to the Global Environmental Facility (GEF) for NGOs*, UNEP-United Nations Foundation.

López Yáñez, Itzamá (2007). *Clasificador Automático de Alto Desempeño* (In Spanish). M.Sc. Thesis. National Polytechnics Institute, Computers Research Center, Mexico

Salazar-Ruiz, E. *et al.* (2008). Development and comparative analysis of tropospheric ozone prediction models using linear and artificial intelligence-based models in Mexicali, Baja California (Mexico) and Calexico, California (US). *Environmental Modelling and Software*, Vol. 23, No. 8, (August 2008) 1056-1069, ISSN: 1364-8152.

Secretaría de Comercio y Fomento Industrial (1986). *Determiación de Neblina de Ácido Fosfórico en los Gases que Fluyen por un Conducto* (in Spanish), Mexican Norm NMX-AA-090-1986, Mexico.

Sistema de Monitoreo Atmosférico de la Ciudad de México (2007). *IMECA* (in Spanish). Available at www.sma.df.gob.mx/simat/pnimeca.htm

Sucar, L.E., Pérez-Brito, J., Ruiz-Suárez, J.C., Morales, E. (1997). Learning Structure from Data and Its Application to Ozone Prediction. *Applied Intelligence*, Vol. 7, No. 4, (November 1997) 327-338, ISSN: 0924-669X.

Toepfer, Klaus *et al.* (2004). *Aliados Naturales: El Programa de las Naciones Unidas para el Medio Ambiente y la sociedad civil* (in Spanish), UNEP-United Nations Foundation.

United Nations (1992). *Rio Declaration on Environment and Development*.

United Nations (1997). *Kyoto Protocol to The United Nations Framework Convention on Climate Change*.

Wang, W., Men, C., Lu, W. (2008). Online prediction model based on support vector machine. *Neurocomputing*, Vol. 71, No. 4-6 (January 2008) 550-558, ISSN: 0925-2312.

Web del Departamento de Medio Ambiente y Vivienda de la Generalitat de Cataluña (in Spanish) (2007). Available at: http://mediambient.gencat.net/cat

Yáñez-Márquez, C., López-Yáñez, I., Sáenz-Morales, G. de la L. (2008). Analysis and Prediction of Air Quality Data with the Gamma Classifier. *Lecture Notes in Computer Science* (ISI Proceedings), LNCS 5197, Springer-Verlag Berlin Heidelberg, 651-658 ISBN: 978-3-540-72394-3.

Pattern Recognition
Edited by Peng-Yeng Yin

ISBN 978-953-307-014-8
Hard cover, 568 pages
Publisher InTech
Published online 01, October, 2009
Published in print edition October, 2009

For more than 40 years, pattern recognition approaches are continuingly improving and have been used in an increasing number of areas with great success. This book discloses recent advances and new ideas in approaches and applications for pattern recognition. The 30 chapters selected in this book cover the major topics in pattern recognition. These chapters propose state-of-the-art approaches and cutting-edge research results. I could not thank enough to the contributions of the authors. This book would not have been possible without their support.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Itzama Lopez-Yanez, Cornelio Yanez-Marquez and Victor Manuel Silva-Garcia (2009). Forecasting Air Quality Data with the Gamma Classifier, Pattern Recognition, Peng-Yeng Yin (Ed.), ISBN: 978-953-307-014-8, InTech, Available from: http://www.intechopen.com/books/pattern-recognition/forecasting-air-quality-data-with-the-gamma-classifier

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821