Pollution, 2(1): 11-23 , Winter 2016
DOI: 10.7508/pj.2016.01.002
Print ISSN 2383-451X Online ISSN: 2383-4501
Web Page: https://jpoll.ut.ac.ir Email: jpoll@ut.ac.ir

A prediction distribution of atmospheric pollutants using support vector machines, discriminant analysis and mapping tools (Case study: Tunisia)

Bedoui, S.[1*], Gomri, S.[2], Samet, H.[1] and Kachouri, A.[1]

1. Research Laboratory on Electronics and information Technologies: LETI National School of Engineers Sfax, University of Sfax, Tunisia
2. Micro Electro Thermal Systems METS Laboratory National School of Engineering of Sfax, University of Sfax, Tunisia

Received: 6 Aug. 2015 Accepted: 18 Sep. 2015

ABSTRACT: Monitoring and controlling air quality parameters form an important subject of atmospheric and environmental research today due to the health impacts caused by the different pollutants present in the urban areas. The support vector machine (SVM), as a supervised learning analysis method, is considered an effective statistical tool for the prediction and analysis of air quality. The work presented here examines the feasibility of applying the SVM to predict the ozone and particle concentrations in two Tunisian cities, namely Tunis and Sfax. We used the SVM with the linear kernel, SVM with the polynomial kernel and SVM with the RBF kernel to predict the ozone and particle concentrations in Tunisia for one year. The RBF kernel produced good results for the two pollutants with 0% error rate. Polynomial and linear kernels produced sufficiently low errors for the pollutants, at 9.09% and 18.18%, respectively. Discriminant Analysis (DA) was selected to analyze the datasets of two air quality parameters, namely ozone O_3 and Suspended Particles SP. The DA results show that the spatial characterization allows for the successful discrimination between the two cities with an error rate of 4.35% in the case of the linear DA and 0% in the case of the quadratic DA. A thematic map of Tunisia was created using the MapInfo software.

Keywords: air pollution, discriminant analysis DA, mapping, ozone, suspended particles, support vector machine SVM

INTRODUCTION

Forecasting of air quality parameters is the common goal for a great number of researches due to the diseases caused by the different gas pollutants. To achieve this goal, support vector machines SVM have been used as statistical tools in air quality prediction and analysis. The SVM model offers a promising alternative and is advantageous in the times series data analysis for predicting the air pollutant levels (Niharika *et al*., 2014). The work by Lu *et al*. (2005) examines the feasibility of applying the support vector machine SVM to predict the air pollutant levels in advancing the time series based on the monitored air pollutant database in the Hong Kong downtown area. Lu *et al*. (2003) presents a pioneer study of using the SVM to forecast the concentration variations of six air pollutants hourly, measured over the duration of the whole year (1999) at the Causeway Bay Roadside Gaseous Monitory Station, one of the fourteen pollutant monitory stations established by the Hong Kong Environment Protection Department (HKEPD) through the Hong Kong territory. Yang *et al*. (2011)

[*] Corresponding Author: souhir.bedoui@yahoo.fr

declares that the SVM can be used for the regression and time series prediction and that it is capable of good generalization while the performance of the model is often hinged on the appropriate choice of the kernel. Several works have been published on the aspect of pollution monitoring using the Discriminant Analysis DA method. In fact, the supervised discriminant analysis method attempts to seek a lower-dimensional space to maximize the separation of the samples from the different classes (Chen *et al*., 2015). Edward *et al*. (2013) used the air quality data collected from eight automatic air quality monitoring stations in central Taiwan and discussed the correlation among the air quality variables with the statistical analysis in an attempt to accurately reflect the differences in air quality observed by each monitoring station, as well as to establish an air quality classification system suitable for the whole of Taiwan. For example, in their study Saithanu *et al*. (2014) utilized cluster analysis to categorize five main pollutants as well as classify the monitoring stations. Discriminant analysis was then constructed to determine the prediction model and evaluate the air quality group. These two statistical multivariate analysis techniques had been applied to assess and predict whether the Air Quality Index AQI in the urban areas in the east of Thailand exceeded the standard level. Discriminant analysis, Cluster analysis and Factor analysis were the aim and objective of Edward *et al*. (2012) to evaluate the Water Quality of a Watershed of Taipei, Taiwan. The use of the Geographical Information System (GIS) to classify polluted regions has also proven to be an effective tool. The scope and scale of the problems in the urban areas make the GIS a powerful tool for the management of spatial and temporal data, complex analyses and visualization (Banja *et al*., 2010). One important GIS application is the mapping of environmental exposure (Bellander *et al*., 2001).

In fact, air pollution is rapidly increasing due to various human activities. It occurs when the environment is contaminated by any chemical, physical or biological agent that modifies the natural characteristics of the atmosphere. Stoves in the homes, vehicles, factories and fires are different sources of air pollution. Both ambient (outdoor) and household (indoor) pollution exert many harmful effects on either human health or the environment. In this work, we focus on ozone and the suspended particulate concentrations in two Tunisian cities.

The physical and chemical processes of the pollutant gases, particularly nitrogen oxides NO_x and the volatile organic compounds (VOC), in the troposphere result in the formation of secondary oxidized products. As several of these processes are regulated by the presence of sunlight, the oxidized products, including an oxidant such as O_3, are commonly referred to as 'secondary photochemical pollutants'. The production of high levels of ground ozone is of particular concern, because it known to act as the primary source of OH and also as a greenhouse gas. Furthermore tropospheric O_3 exerts adverse effects on human health, vegetation and materials (Malec *et al*., 2008).

Suspended particles come from steel, cement, waste incineration and traffic. Particulate matter is linked to major health effects that include ill effects on the breathing and respiratory systems, aggravation of the existing respiratory and cardiovascular diseases, alteration of the body's defense systems against foreign materials, damage to lung tissue, carcinogenesis and premature mortality (Jimoda, 2012).

In this paper, we focus on the concentrations of the ozone and suspended particulates in several Tunisian cities. This paper is organized as follows: Section 2 contains an overview of the atmospheric pollution. It also describes the study area.

A brief description of the support vector machines method, discriminant analysis method and MapInfo software are also included. Section 3 presents the results and discussion of the atmospheric pollutants space distribution using the different measurements with data analysis and mapping methods.

MATERIALS AND METHODS

This section is composed of two parts- first, a brief overview of the ozone and particle sources and the effects. The area under study is presented. The rest of the section is a description of the support vector machines method.

Atmospheric pollution and study area

In Tunisia, the air quality is monitored via fixed and mobile stations. The stations are equipped with various measuring instruments and analyzers of the pollutants such as sulfur dioxide, nitrogen oxides, solid particles, carbon monoxide and ozone. They also include devices for meteorological measurements. The national network for monitoring air quality (RNSQA) was created in 1996 within the National Agency of Environmental Protection (ANPE). It is, in fact, a coherent system with the ability to read the air quality daily in the areas most affected by this phenomenon, like the big cities and industrial zones. Fifteen fixed functional stations and a mobile laboratory are in place.

Support vector machine SVM method

In machine learning, Support Vector Machines (SVM) is a supervised learning model with the associated learning algorithms that analyze the data and recognize the patterns used in the classification and regression analysis (Zhao *et al*., 2013).

SVM identifies the optimal separating hyper-plan between the classes that maximize the margin. Two cases are presented: linearly and non-linearly separable.

Basic steps involved in the SVM algorithm

For the linearly separable case, the hyper plan has Equation (1).

$$f(x) = wx + b = 0 \quad (1)$$

The distance from a point to the plan is:

$$d(x) = \frac{|wx + b|}{\|w\|} \quad (2)$$

Maximize the distance means minimize ||w||.

To minimize ||w||, the α_i the coefficients of Lagrange dual problem should be solved as follows:

$$w = \sum_i \alpha_i x_i y_i \quad (3)$$

$$w^T x + b = \sum_i \alpha_i x_i y_i x^T + b \quad (4)$$

For the non-linearly separable case, kernel functions will be introduced.

Example of such kernels,

- Polynomial kernel (For MATLAB, "d" the order of the polynomial = 3 by default)

$$k(x_i, x) = (x_i x + 1)^d \quad (5)$$

- RBF kernel (δ = 1 by default using MATLAB)

$$k(x_i, x) = \exp\left(-\delta \|x - x_i\|^2\right) \quad (6)$$

Discriminant Analysis (Linear/Quadratic)

The aim is to identify a line in the feature space on which to project all the samples, such that the samples are well (maximally) separated.

Linear Discriminant Analysis LDA is one of the most popular and powerful dimensionality reduction techniques for classification (Wu *et al*., 2015).

LDA provides a linear projection of the data with (c-1) dimensions, taking into account the scatter of data within each class and across the classes. Projection directions are those that maximize the inter-class separation of the projected data (Brahim-Belhouari *et al*., 2005).

Basic steps in the LDA algorithm

Calculate the within class scatter matrix

$$S_W = \sum_{i=1}^{c} S_i, S_i = \sum_i (X_i - \hat{\mu}_i)(X_i - \hat{\mu}_i)^T \quad (7)$$

where $\hat{\mu}_i$ is the mean of each class; and c is the number of classes.

Calculate the between class scatter matrix

$$S_B = \sum_{i=1}^{n} n_i (\hat{\mu}_i - \hat{\mu})(\hat{\mu}_i - \hat{\mu})^T \quad (8)$$

where n_i is the number of observations for each class, $\hat{\mu}_i$ is the mean of each class and $\hat{\mu}$ is the mean of all the classes.

Solve the eigenvalue problem

$$S_B V = \Lambda S_W V \quad (9)$$

The purpose of Fisher-LDA is to maximize the following objective.

$$J(W) = \frac{W^T S_B W}{W^T S_W W} \quad (10)$$

Therefore, the objective of the LDA considers maximizing between the class scatter matrixes and minimizing the within class scatter.

RESULTS AND DISCUSSION

Data base

Ozone O_3

Ozone is monitored in four stations measuring air quality located in District Tunis, Bizerte, Sousse and Sfax cities.

The histogram below illustrates the monthly averages from the ozone registered in the monitoring stations mentioned prior.

From the various histograms (Fig. 1), it becomes clear that the maximum values of the 8-hour averages are recorded during the summer. This is due to the influence of the metrological requirements like temperature and sunshine on the evolution of ozone.

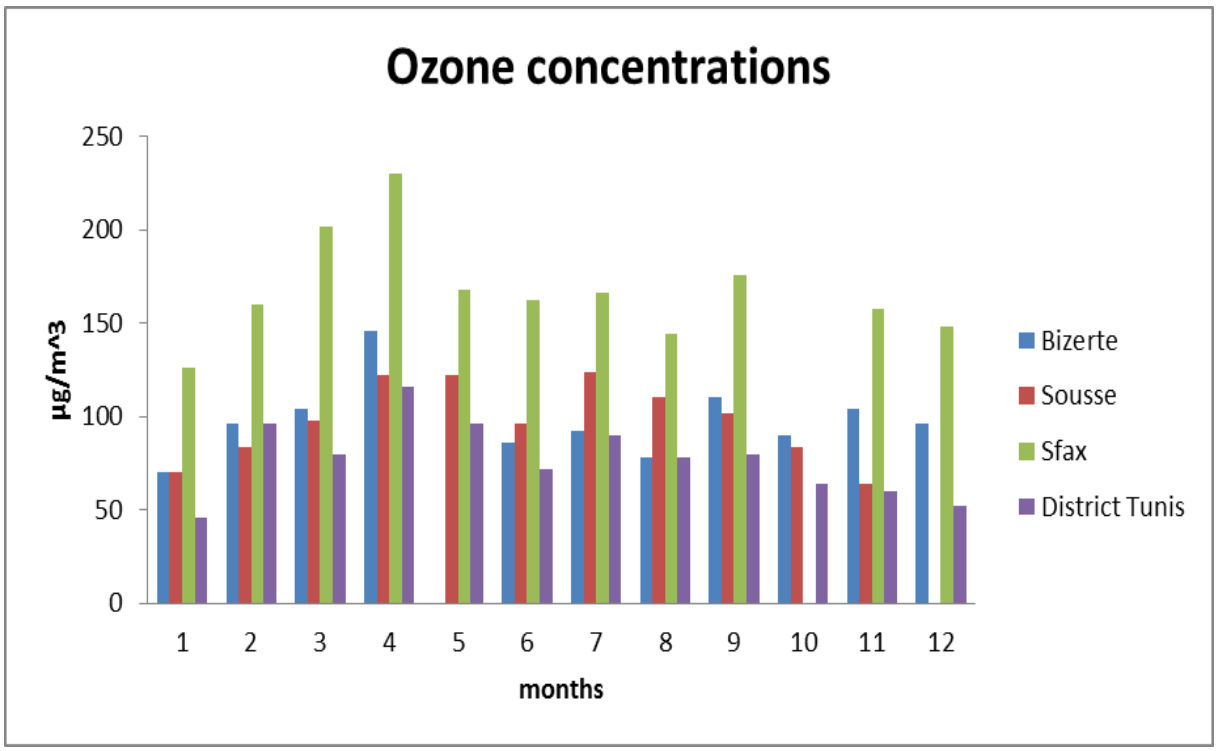

Fig. 1. Maximum monthly ozone concentrations in four cities

Suspended particles

We present via the graph given above, the evolution of the monthly average of the Suspended particles SP monitored in the four stations measuring the air quality located in District Tunis, Bizerte, Sousse and Sfax cities.

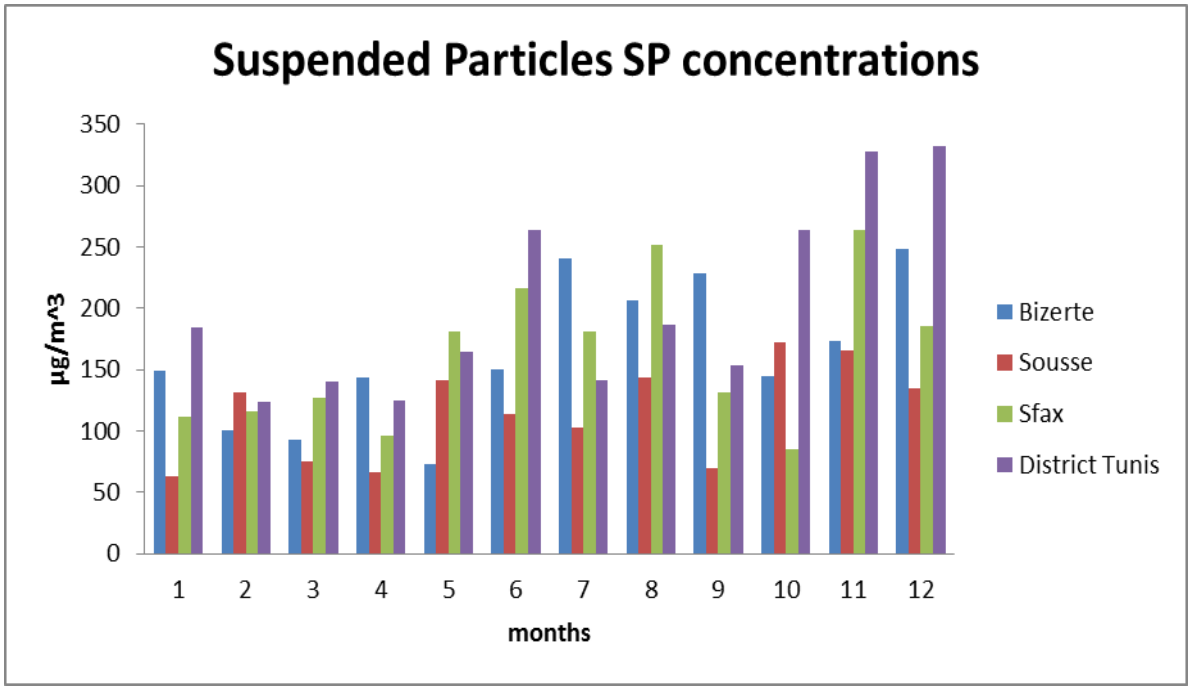

Fig. 2. Maximum monthly Suspended Particles SP concentrations in four cities

From the histograms (Fig. 2), we demonstrate that the maximum concentrations were recorded during the months of October, November and December for the various stations. The variations in the concentration of the suspended particles also follow a seasonal classic variation: the highest levels are observed during the winter season (October-February) against the summer season (March-September) which is characterized by the presence of low stable concentrations.

Gas dispersion with the SVM method

The monthly concentrations of the ozone and particles for each city were used as the input parameters for the SVM. We explored the discrimination between the two cities (District Tunis/Sfax). For the purpose of classification, the data were divided into two subsets; training and testing. In fact, separating the data into training and testing sets is essential to evaluate the data in the support vector machine model. Using MATLAB we applied the SVM to our database. The function "svmtrain" is used for learning an SVM classifier. The following instructions are used in the simulation:

Svmstruct = svmtrain (data (train,), groups (train) 'KERNEL_FUNCTION', 'Kfun', 'show plot' true) with Kfun it can be ‘linear’, 'rbf' or 'polynomial'.

The results for the linear SVM kernel, RBF SVM kernel and Polynomial kernel are revealed in Figures 3, 4 and 5, respectively.

Table 1 evaluates the performance of the classifier.

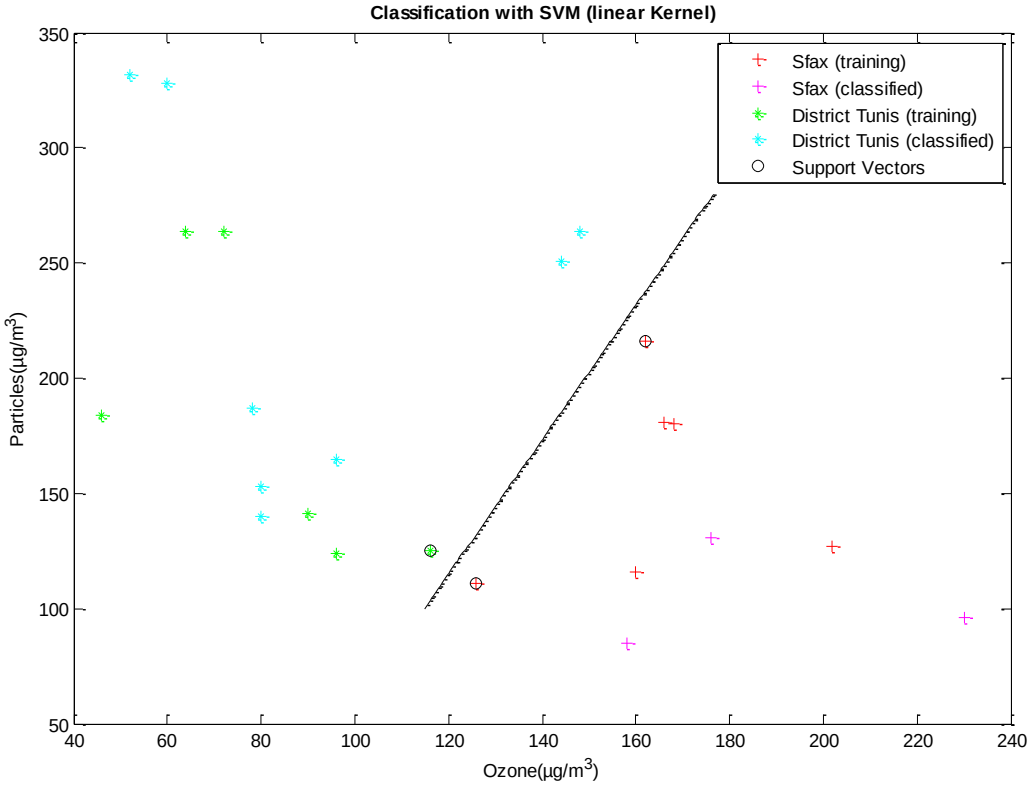

Fig. 3. SVM linear kernel for District Tunis and Sfax

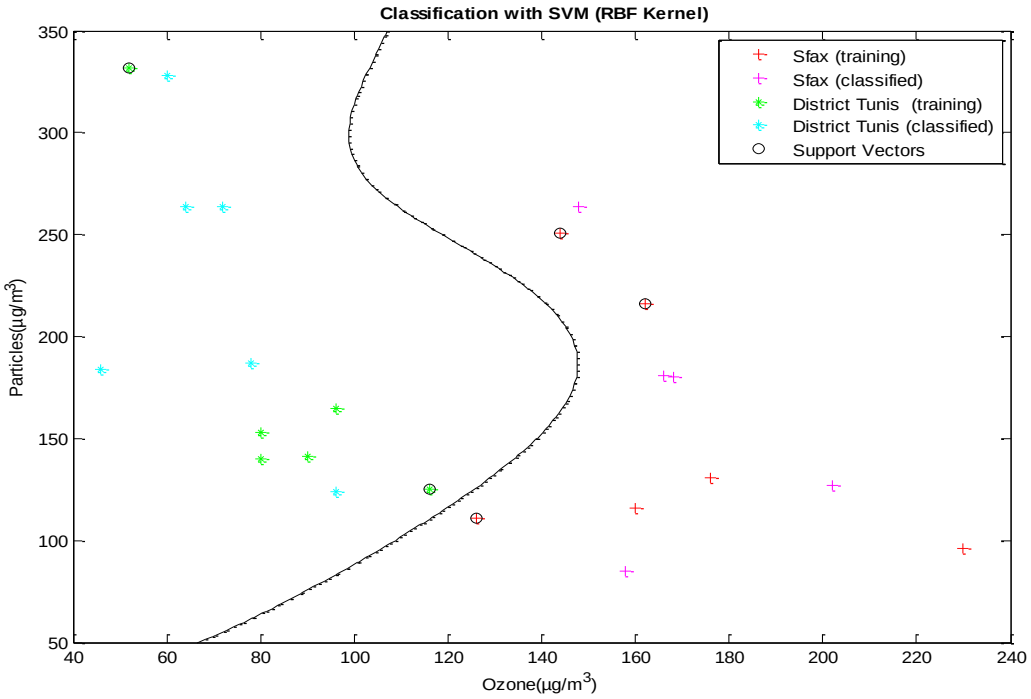

Fig. 4. SVM RBF kernel for District Tunis and Sfax

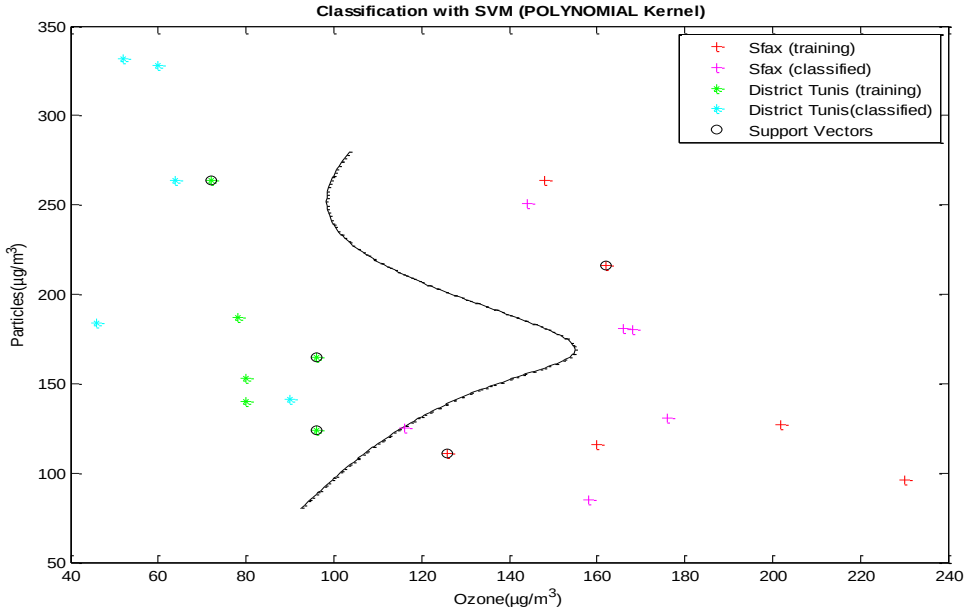

Fig. 5. SVM Polynomial kernel for District Tunis and Sfax

Table 1. Summary of the performance of different kernels used in the SVM analysis

			Correct rate	Error rate	Sensitivity	Specificity
Linear kernel			81.82%	18.18%	0.8	1
Rbf kernel	Sigma	0.1	54.55%	45.45%	0	1
		0.5	81.82%	18.18%	0.6	1
		1	100%	0%	1	0.8333
polynomial kernel	Order	2	81.82%	18.18%	0.6	1
		3	90.91%	9.09%	0.8	1

The evaluation of the performance in the support vector classification was based on the sensitivity, specificity, correct rate and error rate of the forecast. The sensitivity, specificity, correct and error rate are calculated based on the Equations (7), (8), (9) and (10).

$$\text{Correct rate} = \frac{\text{correctly classified samples}}{\text{classified samples}} \tag{11}$$

$$\text{Error rate} = \frac{\text{incorrectly classified samples}}{\text{classified samples}} \tag{12}$$

$$\text{Sensitivity} = \frac{\text{correctly classified positivesamples}}{\text{true positivesamples}} \tag{13}$$

$$\text{Specificity} = \frac{\text{correctly classified negative samples}}{\text{true negative samples}} \tag{14}$$

From the Table 1, we observe that the SVM is capable of good prediction while the performance of the SVM model is often dependent on the kernel selection. Among the three kernels that have been used to train the data, the RBF kernel has been selected because of its high values and the correct rate. The appropriate choice of the

kernel RBF kernel produced good results for the two pollutants. The polynomial and linear kernels produced adequate but low errors for the pollutants.

Gas dispersion with DA method

The monthly concentrations of the ozone and particles for each city were used as the input parameters for the Discriminant analysis. In the first step, we study the discrimination among four cities (District Tunis/Bizerte/Sousse/ Sfax). The data includes measurements on the ozone emission and the particle emission of all four cities. There are twelve measurements for Tunis city, and eleven measurements for the other three cities. We loaded the data and observed the differences in the gas emissions among the cities. If we measured the ozone and particles of a city and needed to determine its membership based on those measurements, one approach towards solving this problem could be used, termed discriminant analysis. The classify function can perform the classification using the different types of discriminant analysis. First, we classify the data using the default linear method as shown in Figure 6.

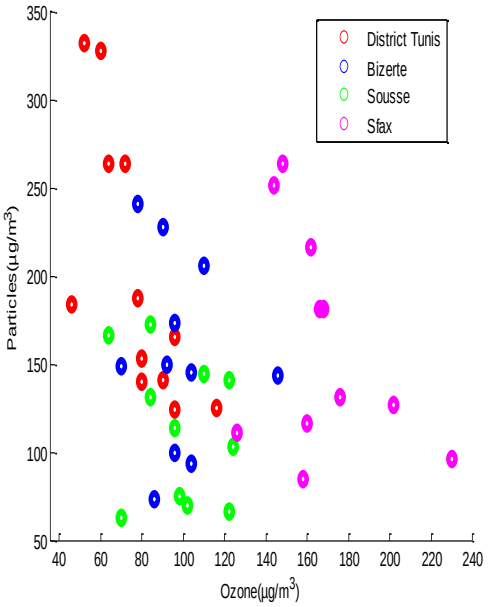

Fig. 6. Classification with linear method

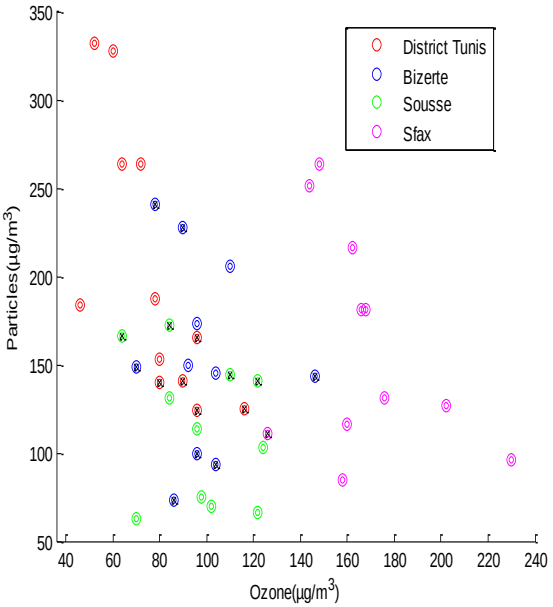

Fig. 7. Misclassified points with linear method

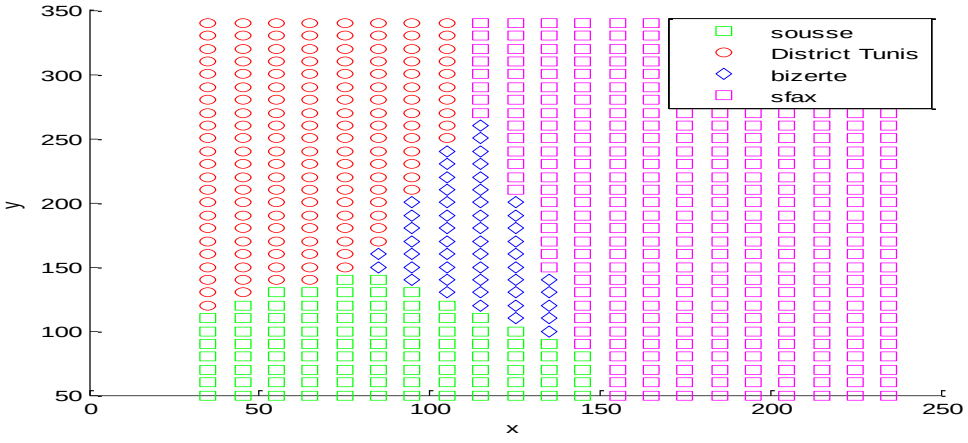

Fig. 8. Separated regions with linear method

Among the 45 measurements, 37.78% or 17 values are misclassified by the linear discriminant function. We can identify those by drawing an X through the misclassified points (Fig. 7). This function separates the plane into four regions divided by lines, and different regions have been assigned to different species. One way to visualize these regions is to create a grid of (x, y) values and apply the classification function to that grid, as illustrated in Figure 8. For some data sets, the regions for the various groups are not clearly separated by lines. In such instances, the linear discriminant analysis is inadequate. Therefore, the proportion of the misclassified points can be re-computed using the quadratic discriminant analysis. In the case of four classes, the quadratic discriminant analysis also proves inadequate. Each method misclassifies 37.78% of the specimens. In fact, 37.78% may be an underestimate of the proportion of misclassified items that could be expected if we classified a new data set.

In this section, we study the discrimination between the two cities (District Tunis/ Sfax). The data consists of measurements on the ozone emission and the particle emission of the two cities. Twelve measurements are given for Tunis city and eleven measurements for Sfax city. After loading the data we observed the differences between the emissions of gases between the cities. The classify function can perform the classification using different types of discriminant analysis. First, we classified the data using the default linear method, as shown in Figure 9.

Among the 23 measurements, 4.35% or 1 value is misclassified by the linear discriminant function. We can identify them by drawing an X through the misclassified points as in Figure 10 shown below. In fact, the function separates the plane into two regions divided by lines, and assigns each region to a city. We created a grid of (*x* and *y*) values and applied the classification function to that grid to visualize these regions (Fig. 11). For some data sets, the regions for the various groups do not get clearly separated by lines. When this is the case, using the linear discriminant analysis is not appropriate. We can recalculate the proportion of the misclassified observations using the quadratic discriminant analysis. In this case, of the 23 measurements, 0% or no value is misclassified by the quadratic discriminant function as shown in Figure 12.

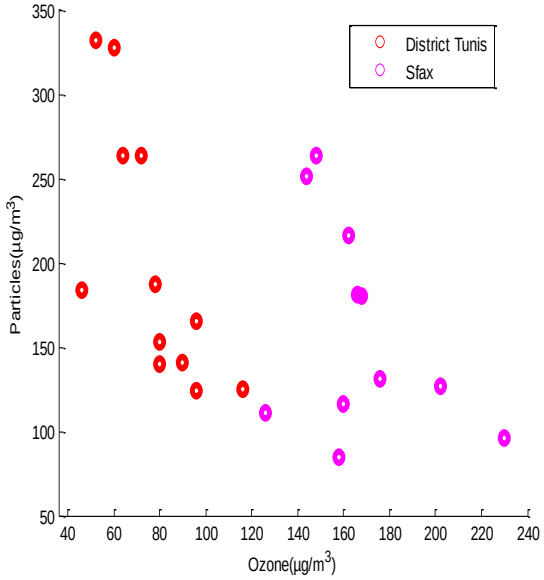

Fig. 9. Classification with linear method

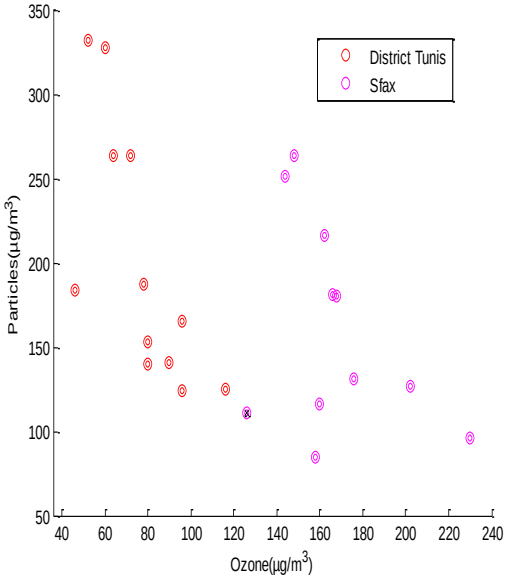

Fig. 10. Misclassified points with linear method

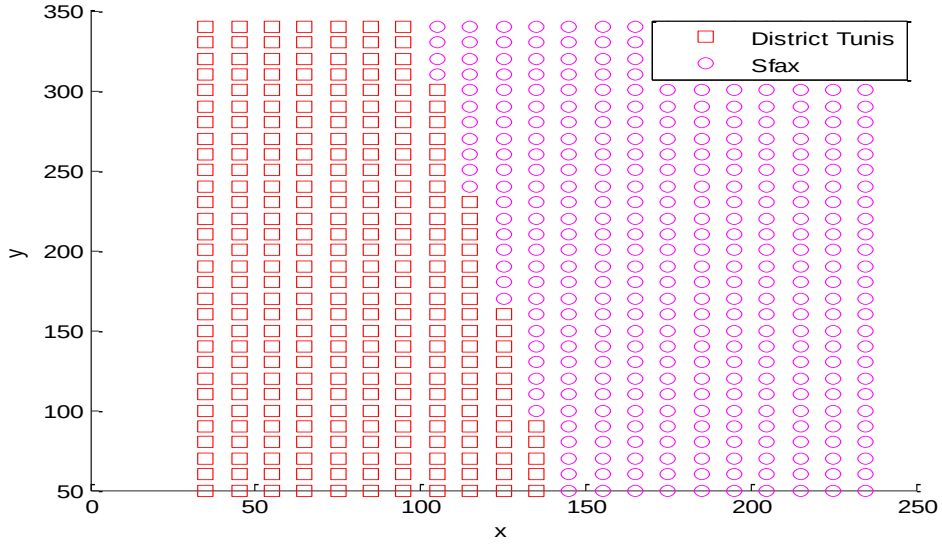

Fig. 11. Separated regions with linear method

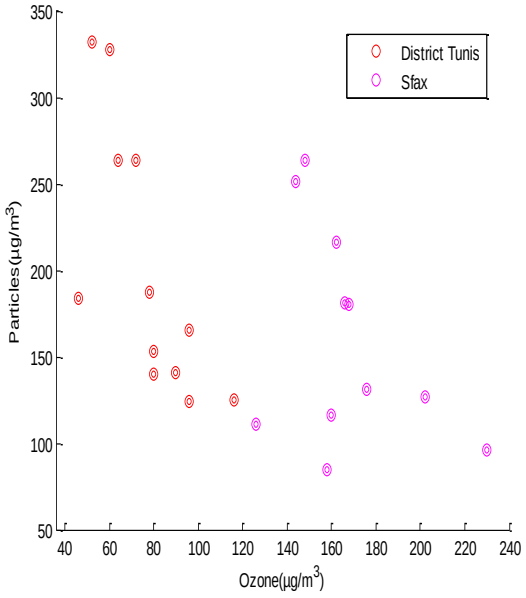

Fig. 12. Misclassified point with the quadratic method

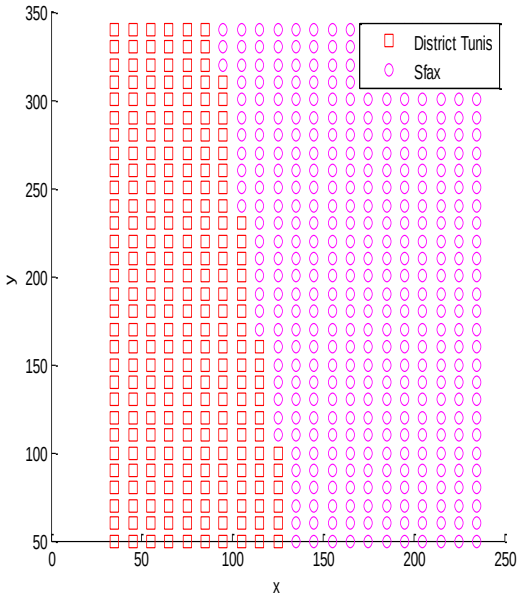

Fig. 13. Separated regions with the quadratic method

The function has separated the plane into two regions divided by lines, and each region is assigned to a city. A grid of *x* and *y* values the classification function so that the grid was created to visualize these two regions in Figure 13 given below.

Table 2. Comparison of results obtained from a linear and quadratic discriminant analysis

		Data size	Error rate	Correct rate
4 classes	Linear DA	45	37.78%	62.22%
	Quadratic DA	45	37.78%	62.22%
2 classes	Linear DA	23	4.35 %	95.65%
	Quadratic DA	23	0%	100%

Table 2 is a summary table. It recapitulates the different results. In the case of four classes either assigned linear DA or quadratic DA does not perform gaseous dispersion among the regions. These two methods produce 37.78% misclassified rate. In the case of two classes, the quadratic DA produces a lower misallocation rate than does the linear DA.

Gas dispersion with MapInfo

MapInfo software is a Geographical Information System (GIS) widely used in the field of the environment. We plotted each average on a Tunisia map according to the regions using the Geographical Information System software (MapInfo 11.5).

To create a thematic map with MapInfo we follow the steps presented in Figure 14.

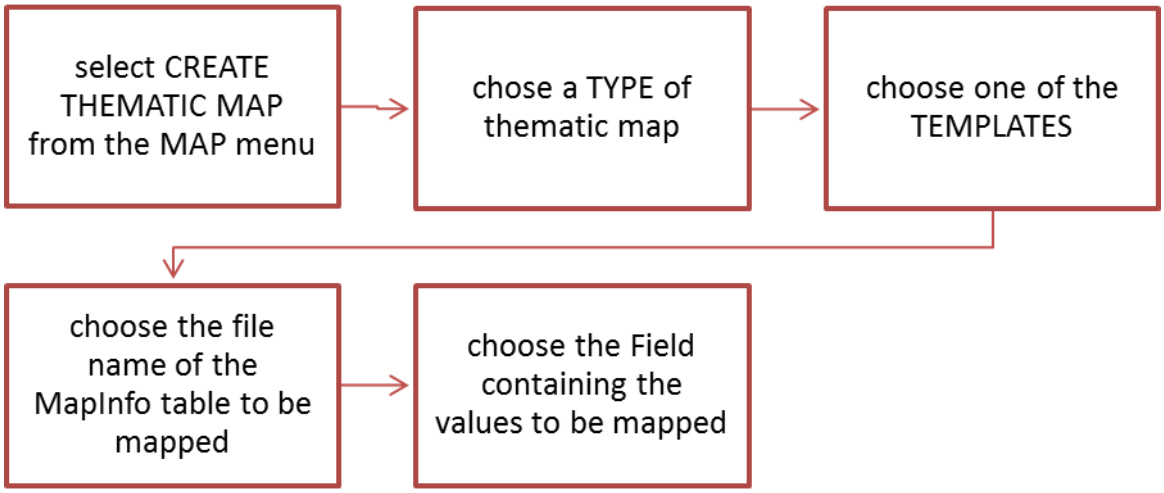

Fig. 14. Thematic map with MapInfo software steps

The result is shown in Figure 15.

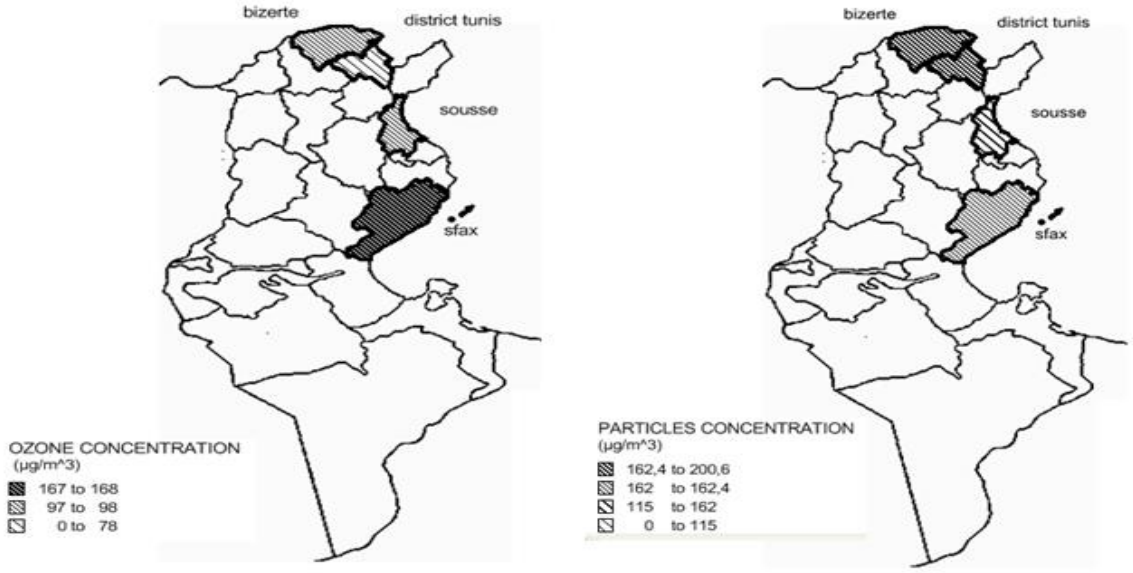

Fig. 15. Annually the Ozone and Suspended particle (SP) concentrations in Tunis, Bizerte, Sousse and Sfax regions

From the map shown above (Fig. 15) it is very easy to understand the information. The process used to represent the gas concentration is a graduation of color. The basic color of the graduation selected is black. Four regions namely Bizerte, District Tunis, Sousse and Sfax have been classified using the MapInfo software. In fact, thematic maps improve the understanding of the data. The user is able to actually see the polluted states of the cities, to analyze and to compare the situations in different regions. It is true Sfax has the higher ozone concentrations compared with the other regions. The maximum concentrations for the Suspended particles SP were detected at Tunis.

CONCLUSION

In this paper an analysis of the dispersion of ozone and the suspended particles is presented.

Controlling and forecasting the air quality parameters have become an important subject in atmospheric and environmental research due to the health and environmental effects caused by exposures to the air pollutants in the urban areas. In this paper, Support Vector Machines SVM has been performed to predict the concentrations of the ozone and particles. To achieve this goal, the SVM with the linear kernel, SVM with the polynomial kernel and SVM with the RBF kernel have been used.

For future work, more SVM kernels can be implemented and comparisons can be performed to identify which kernels give better prediction. The concentrations of these pollutants have also been investigated by a discriminant analysis DA of the monthly average values measured at four, and then at two locations in Tunisia. In the case of four classes, regardless of either linear or quadratic no reliable classification is available, with an error rate of 37.78%. However, in the case of two classes, the quadratic DA produces a lower misallocation rate (0%) than does the linear DA (4.35%). A thematic map was created using the MapInfo software enabling classification of the regions.

ACKNOWLEDGEMENTS

The authors extend their thanks to the Tunisian National Institute of Meteorology for providing the different values for the gas pollutants.

REFERENCES

Banja, M., Como, E., Murtaj, B. and Zotaj, A. (2010). Mapping air pollution in urban tirana area using GIS. (Paper presented at International Conference SDI 2010 – Skopje)

Bellander, T., Berglind, N., Gustavsson, P., Jonson, T., Nyberg, F., Pershagen G. and Järup, L. (2001). Using Geographic Information Systems to Assess Individual Historical Exposure to Air Pollution from Traffic and House Heating in Stockholm. Environ Health Persp, 109(6), 633-639.

Brahim-Belhouari, S. and Bermak, A. (2005). Gas identification using density models" Pattern Recogn Lett, 26, 699–706.

Chen, C., Zhang, Z., Ouyang, M.,Liu, X.,Yi, L.,Liang, Y. and Zhang, C. (2015). Shrunken centroids regularized discriminant analysis as a promising strategy for metabolomics data exploration. J. Chemometrics, 29(3), 154-164.

Edward, M.Y and Kuo, S.L. (2012). Applying a Multivariate Statistical Analysis Model to Evaluate the Water Quality of a Watershed. Water Environ Res, 84(12), 2075-2085.

Edward, M.Y and Kuo, S.L. (2013). A Study on the Use of a Statistical Analysis Model to Monitor Air Pollution Status in an Air Quality Total Quantity Control District. Atmosphere, 4(4), 349-364.

Jimoda, L. A. (2012). Effects of particulate matter on human health, the ecosystem, climate and materials: a review. FU,,Work Liv Env Prot, 9(1), 27-44.

Lu, W.Z. and Wang, W.J. (2005). Potential assessment of the support vector machine method in forecasting ambient air pollutant trend. Chemosphere, 59(5), 693–701.

Lu, W., Wang, W., Wang, X. and Leung, A.Y.T. (2003). Prediction of Air Pollutant Levels using Support Vector Machines: An Effective Tool. (in B.H.V. Topping, (Editor), "Proceedings of the Seventh International Conference on the Application of Artificial Intelligence to Civil and Structural Engineering", Civil-Comp Press,

Stirlingshire, UK, Paper 52, 2003. doi:10.4203/ccp.78.52)

Malec, L. and Skacel, F. (2008). Analyzing ground ozone formation regimes using a principal axis factoring method: A case study of Kladno Czech Republic indistrual area. Atmosfera, 21(3), 249-263.

Niharika., Venkatadri, M. and Padma, S. (2014). A survey on Air Quality forecasting Techniques. Int. j. comput. sci. inf. technol, 5(1), 103-107.

Wu, G. and Feng, T. (2015). A theoretical contribution to the fast implementation of null linear discriminant analysis with random matrix multiplication. Numer. Linear Algebra Appl, doi:10.1002/nla.1990.

Yang, J.Y., Ip, W.F., Vong, C.M. and Wong, P.K. (2011, June). Effect of Choice of Kernel in Support Vector Machines on Ambient Air Pollution Forecasting. (Paper presented at International Conference on System Science and Engineering, Macau, China)

Saithanu, K. and Mekparyup, J. (2014). Air quality assessment in the urban areas with multivariate statitical analysis at the east of Thailand. Int J Pure Appl Math, 91(2), 169-177.

Zhao, G., Song, J. and Song, J. (2013). Analysis about Performance of Multiclass SVM Applying in IDS. (Paper presented at International Conference on Information, Business and Education Technology ICIBIT)