International Journal of Pure and Applied Mathematics
Volume 84 No. 2 2013, 109-121
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
doi: http://dx.doi.org/10.12732/ijpam.v84i2.9
ijpam.eu

ASSESSMENT AND PREDICTION OF THE GROUND LEVEL OZONE CONCENTRATION IN THE EAST OF THAILAND

Kidakan Saithanu[1], Jatupat Mekparyup[2 §]
[1,2]Department of Mathematics
Burapha University
169, Tambon Saensook, Amphur Muang, Chonburi, 20131, THAILAND

Abstract: To assess and predict whether the ground level ozone concentration exceeds an air quality standard in ambient, two different techniques have been applied. One is the traditional method, discriminant analysis model, and the other is an alternative scheme, neural network model. Daily ground ozone maximum concentration and other diverse variables in the air, measured from the monitoring stations in the east of Thailand for the period 2006-2010, were used to train and validate these two predictive models. The performance of the models can be evaluated by a correct classification rate (CCR). The result of performance comparison indicates the neural network model is shown to overcome the classical discriminant analysis model for both the training and the validation data set. That is, the average CCR of the neural network model is 87.22% for the training data set and 86.58% for the validation data set while the average CCR of the discriminant analysis model provides 79.77% and 78.98% for the training and the validation data set, respectively.

AMS Subject Classification: 62H30
Key Words: ground level ozone, factor analysis, discriminant analysis, neural network

1. Introduction

The tropospheric ozone or ground level ozone (O_3) has harmful effects on hu-

Received: February 16, 2013

© 2013 Academic Publications, Ltd.
url: www.acadpubl.eu

[§] Correspondence author

man health, vegetation and ecosystems. The ground level ozone concentration depends on not only a sophisticated chemical process but also a combination of local meteorological conditions. Ozone formation is formed as a result of a series of complex chemical reaction of Volatile Organic Compounds (VOCs) and Nitrogen Oxides (NO_X) in the presence of heat and sunlight. VOCs and NO_X are emitted from a variety of sources, including motor vehicles, chemical or power plants, refineries, factories, other industrial sources, and other combustion devices. Ozone concentration is also strongly linked to meteorological conditions. In addition, favorable conditions (warm temperature, soft wind, land breeze or sea breeze at a coastal site) have a great influence on ozone concentration [1], [2]. Therefore, ozone concentration is currently a matter of large concern. Furthermore, the ground level ozone is one of five main pollutants (PM_{10}, SO_2, CO, NO_2 and O_3) which is monitored and pointed the Air Quality Index (AQI). Each of these pollutants has an air quality standard which is used to calculate the overall AQI. Simultaneously, one can also establish the limiting pollutant(s), resulting in the estimating AQI. In number, AQI is explained from 0 to 500 with 0 explaining good air and 500 explaining hazardous air [3].

Due to Chonburi and Rayong are the industrial centers and urban coastal areas in the east of Thailand, many serious air pollution problems have been increasing including ozone pollution. According to AQI stipulated by the Thai Environmental Protection Department, the standard of 1 hour ozone level in the air is 100 part per billion (ppb.) or 200 $\mu g/m^3$ [4]. The annual concentration report of Chonburi [5] as well as the air quality report of Rayong [6] informed ozone concentrations were more frequently surpassed the standard level. Hence, an accurate ozone alert forecasting is necessary to issue warnings to the public before the ozone concentration reaches a dangerous level.

The purposes of this work are two-fold: (1) to determine significant variables that have influential impact on ozone concentration. (2) to obtain predictive models as guidance tools to predict and classify observations of ozone concentrations into three groups of air quality corresponding to the AQI of specific ozone pollutant. Each of these air quality groups would be represented by ozone concentrations of: 0-50 ppb., for the good air, 51-100 ppb., for the moderate air, and >100 ppb., for the unhealthy air for sensitive people.

Since factor analysis is a data reduction technique, it is an appropriate statistical method in determining the important air quality and meteorological variables underlying ground ozone formation such as [7]. Discriminant analysis and neural network models were proposed techniques for classification of ozone concentration. Discriminant analysis is a conventional method based on statistical assumptions so it is widely used in general as of [7], [8], [9]. Unlike other statistical methods, neural network makes no prior assumptions concerning the data distribution. Many papers [10], [11], [12], [13], [14] applied the multilayer perceptron (MLP) neural network model to investigate the importance of local air quality and meteorology variables in categorizing the groups of ozone concentrations.

2. Data Description

The General Education Centre, Mueang District, Chonburi and the Map Ta Phut Health Office, Mueang District, Rayong were the representatives of the eastern monitoring stations in Thailand to measure the daily ground ozone maximum concentration and both of air quality and meteorological variables for the period 2006-2010. The daily ground ozone maximum concentration was considered as the dependent variable which was categorized into three groups of air quality: good, moderate and unhealthy. The following 16 variables regarded as the independent variables were following: concentrations of nine air quality variables (CO, NO, NO_2, NO_X, SO_2, HC, CH_4, NMHC and PM_{10}) and seven meteorological variables (pressure, rain, relative humidity (RH), temperature (Temp), sun radiation (SR), wind direction (WD) and wind speed (WS)). For this analysis, two mutually exclusive and distinct data sets were created to train and validate the two predictive models, discriminant analysis and neural network models. The training data set (1,542 cases) measured for the period 2006-2009 was used to develop classifiers. The validation data set (723 cases) measured in 2010 was considered to determine how well each classifier would perform against a data set that was not used in the training data set.

3. Methodology

3.1. Determination of Significant Variables

Factor analysis is useful method to explain correlations among observable variables in terms of unobservable variables called factors. For a numerical realization and under certain conditions, the sample correlation matrix (**R**) can be expressed as Equation 1 [15].

$$\mathbf{R} = \mathbf{L}\mathbf{L}' + \psi \tag{1}$$

where $\mathbf{L}$ is the matrix of factor loadings of the ith variable on the jth factor (l_{ij}) and ψ is the matrix of specific variances. The quantity $h_i^2 = 1 - \psi_i$ is defined as the communality. In other words, the communality of any variables is less than or equal to the reliability of that variable. The factoring estimator l_{ij} is obtained by maximizing the variance of a factor to all variables using an iterative algorithm with two of the most popular methods for parameter estimation, the principal component method and the maximum likelihood method. The solution from either method can be rotated with varimax rotation in order to gain a clear association of factors to the original variables. Generally, factor analysis is centered on the parameter in the factor model. However, the estimated values of the common factors, called factor scores, may also be required. The calculation of factor scores $(\mathbf{F})$ is displayed as Equation 2 [15].

$$\mathbf{F} = \left(\mathbf{L}\mathbf{L}^{\prime}\right)^{-1}\mathbf{L}^{\prime}\mathbf{X} \tag{2}$$

where $\mathbf{X}$ is the vector of observations and the term of matrix $(\mathbf{L}\mathbf{L}^{\prime})^{-1}\mathbf{L}^{\prime}$ may be called the coefficient matrix. Factor scores are the quantities often used for diagnostic objectives, as well as inputs to a subsequent analysis.

For this study, there are numerous variables in the air which are related to each other so factor analysis is utilized to determine the significant variables which are crucial to ozone concentration. Factor analysis seeks to find connections among the air quality variables (CO, NO, NO_2, NO_X, SO_2, HC, CH_4, NMHC and PM_{10}) or the meteorological variables (Pressure, Rain, RH, Temp, SR, WD and WS) by finding an explanation for the values in the sample correlation matrix.

3.2. Discriminant Analysis Model

Discriminant analysis is a pattern recognition technique. It is concerned with separating distinct sets of objects (or observations) and with allocating new objects to previously defined groups. The task of this analysis tries to find the combination of variables that best predicts the category or group to which a case belongs. The group identification must be known for each case used in the analysis. The combination of predictor variables is called a classification function, and then this function can be used to classify new cases whose group membership is unknown. Two ways of making discriminant analysis are linear and quadratic discriminant analysis. On linear discrimination technology, the covariance matrices are assumed equal for all groups. An estimate $(\hat{d}_i(\mathbf{x}))$ of the linear discriminant score $(d_i(\mathbf{x}))$ is defined as Equation 3 [15].

$$\hat{d}_i\left(\mathbf{x}\right) = \overline{\mathbf{x}}_i^{\square}\mathbf{S}_{pooled}^{-1}\mathbf{x} - \frac{1}{2}\overline{\mathbf{x}}_i^{\square}\mathbf{S}_{pooled}^{-1}\overline{\mathbf{x}}_i + \ln\left(p_i\right) \tag{3}$$

where $\overline{\mathbf{x}}_i$ is the vector of mean for group $i \; ; i = 1, 2, \ldots, g$, $\mathbf{S}_{pooled}$ is the pooled variance matrix and p_i is the prior probability. Then $\mathbf{x}$ would be allocated to population π_k if $\hat{d}_i\left(\mathbf{x}\right)$ is the largest of $\hat{d}_1\left(\mathbf{x}\right), \hat{d}_2\left(\mathbf{x}\right), \ldots, \hat{d}_g\left(\mathbf{x}\right)$; $k \neq i$.

On quadratic discrimination technology, the assumption of homogeneity of within groups is patently violated, although it retains the distributional assumption of multivariate normality. An estimate $(\hat{d}_i^Q(\mathbf{x}))$ of quadratic discriminant score $(d_i^Q(\mathbf{x}))$ is defined as Equation 4 [15].

$$\hat{d}_i^Q\left(\mathbf{x}\right) = -\frac{1}{2}\ln\left(\left|\mathbf{S}_i\right|\right) - \frac{1}{2}\left(\mathbf{x} - \overline{\mathbf{x}}_i\right)^{\square}\mathbf{S}_i^{-1}\left(\mathbf{x} - \overline{\mathbf{x}}_i\right) + \ln\left(p_i\right) \tag{4}$$

where $\mathbf{S}_i$ is the covariance matrix for group $i \; ; i = 1, 2, \ldots, g$. Then $\mathbf{x}$ would be allocated to population π_k if $\hat{d}_i^Q\left(\mathbf{x}\right)$ is the largest of $\hat{d}_1^Q\left(\mathbf{x}\right), \hat{d}_2^Q\left(\mathbf{x}\right), \ldots, \hat{d}_g^Q\left(\mathbf{x}\right)$; $k \neq i$.

3.3. Neural Network Model

The neural network approach does not require any assumptions. It only finds relationships between independent variables (input variables) and dependent variables (output variables) from existing data. Thus, an appropriately designed neural network may be easier for practitioners to use and more robust to customary methods. The neural network is often described as layers of functional nodes or neurons. Nodes of the neural network are connected by the weights. The most common and popular neural network architecture is the feed-forward MLP which typically contains 3 layers. The first layer consists simply of the input variables, all of which are connected to a hidden layer which is the second layer. Finally, the third layer is the output layer. The computation occurs at the nodes in both the hidden and the output layers. The input layer only passes the data to other layers. The function that is used at the computing nodes is called the activation function or transfer function. There is no rule of thumb to determine the number of hidden layers or the number of nodes in each hidden layer. Hence, the architecture feature of neural network depends on crucial factors such as number of hidden layers, number of hidden layer nodes, number of output nodes and the type of transfer function. However, the neural network model can solve a problem by finding optimal weights.

A full feed-forward MLP network with backpropagation was considered for this study because of its simplicity. [16], [17] recommended one hidden layer network is sufficient to model any complicated problem so the designed network model would only have one hidden layer. As [18] reported, there is no standard rule for designing the optimum number of hidden layer nodes. Few nodes may not be enough for the neural network to provide adequate model due to undertraining. Alternatively, overtraining happens if there are too many nodes which increase the size of network. Thus, this study was designed to have one hidden layer with 3 nodes. The number of input layer nodes is generally associated with the number of observations that are required to represent each set of input variables. Finally, the number of output layer nodes is related to the expected number of different classifications which need to be determined by neural network.

3.4. Performance Criterion

There can be many criterions to measure and compare the performance of model. A correct classification rate (CCR) is the most one frequently used [19], [20]. It is defined as Equation 5.

$$CCR = \frac{\sum\limits_{k=0}^{C-1} CC_k}{n} \tag{5}$$

where CC_k is the number of correctly classified observations and n is the number of observations in the class. The model with the highest correct classification rate is the one with the better performance. Generally, the CCR is applied to judge the functional network classifier performance. The better classifier is the one with the highest CCR.

4. Results

4.1. Results of Determination of Significant Variables

A suitable number of factors are to set the number of eigenvalues greater than 1 or nearly to 1. As of Table 1, the estimates of the varimax rotation factor loadings with the principal component method for the air quality variables indicated that all the oxide of nitrogen variables (NO, NO_2 and NO_X) loaded positive highly on the first factor (F_1) so it might be called the OxideNitrogen

Factor	Eigenvalue	Air Quality Variables	Factor Loadings
F_1	2.9046	NO	0.838
		NO_2	0.650
		NO_X	0.923
F_2	1.7302	CO	0.808
		PM_{10}	0.816
F_3	1.3731	HC	−0.773
		CH_4	−0.930
F_4	1.0647	NMHC	0.979
F_5	0.8497	SO_2	0.966

Table 1: Varimax rotation factor loadings for the air quality variables with the principal component method

factor. Similarly, the positive high loadings of CO and PM_{10} variables in the second factor could be called the CO&PM_{10} factor. The hydro carbon variables (HC and CH_4) had relatively large negative loadings on the third factor (F_3) then they would be HCGroup factor. The remaining factors might be explained as the fourth factor (F_4) appeared to be primarily a NMHC factor and the fifth factor (F_5) labeled with a SO_2 factor.

Table 2 showed the estimates of varimax rotation factor loadings with the maximum likelihood method for the meteorological variables. The sun radiation loaded highly and the temperature loaded moderately on the first factor so they would be Temp&SR factor. The pressure was the only variable with a large loading on the second factor thus it might be a pressure factor. Rain, relative humidity, wind direction and wind speed variables loaded relatively negative loadings on the third factor so it might be called the Humid&Wind factor.

Next, the inputs or the important independent variables in determination of ozone concentration were calculated by the coefficient matrix in Equation 2. Finally the eight independent variables (OxideNitrogen, CO&PM_{10}, HCGroup, NMHC, SO_2, Temp&SR, Pressure and Humid&Wind) were trained and validated in both of the discriminant analysis and the neural network models.

4.2. Results of Discriminant Analysis Model

Since the covariance matrices of all air quality groups were equal, the linear discrimination technology was applied. Once the eight significant variables were evaluated by discriminant analysis, the linear discriminant score for each

Factor	Eigenvalue	Meteorological Variables	Factor Loadings
F_1	1.4342	Temp	0.577
		SR	0.796
F_2	1.0635	Pressure	0.997
F_3	0.8802	Rain	−0.098
		RH	−0.901
		WD	−0.351
		WS	−0.036

Table 2: Varimax rotation factor loadings for the meteorological variables with the maximum likelihood method

of air quality groups was below:

For group of good air:

$$\begin{aligned}\hat{Y}_0 = -64.363 - 0.464 OxideNitrogen + 1.079 CO\&PM_{10} - 12.229 HCGroup \\ - 4.629 NMHC + 0.253 SO_2 - 0.135 Temp\&SR + 0.165\,\mathrm{Pr}\,essure \\ - 0.440 Humid\&Wind\end{aligned}$$

For group of moderate air:

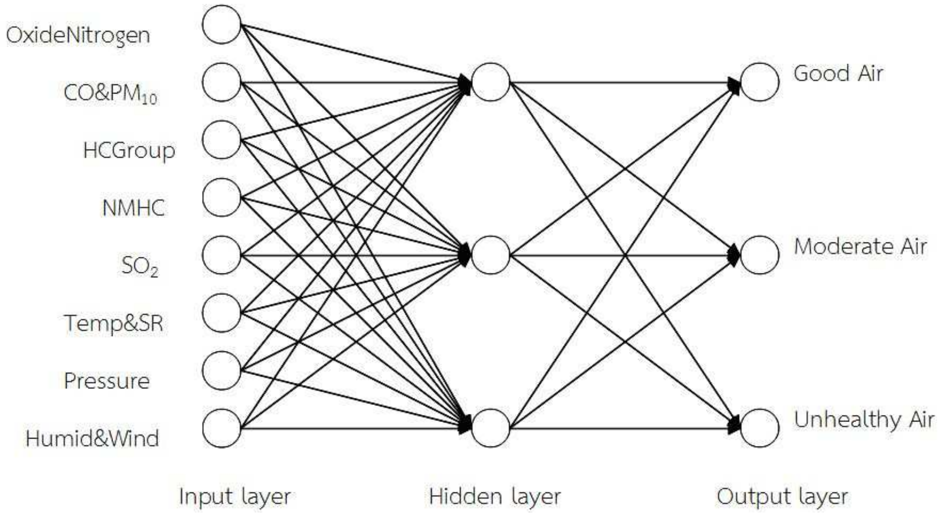
$$\begin{aligned}\hat{Y}_1 = -60.182 - 0.490 OxideNitrogen + 1.518 CO\&PM_{10} - 13.753 HCGroup \\ - 4.385 NMHC + 0.369 SO_2 - 0.104 Temp\&SR + 0.143\,\mathrm{Pr}\,essure \\ - 0.362 Humid\&Wind\end{aligned}$$

For group of unhealthy air:

$$\begin{aligned}\hat{Y}_2 = -75.475 - 0.427 OxideNitrogen + 2.172 CO\&PM_{10} - 16.668 HCGroup \\ - 4.257 NMHC + 0.493 SO_2 - 0.119 Temp\&SR + 0.152\,\mathrm{Pr}\,essure \\ - 0.410 Humid\&Wind\end{aligned}$$

4.3. Results of Neural Network Model

Once the neural network model was designed according to the results of significant variables associated with ozone concentration, the input layer consisted of 8 nodes. In addition, the output layer nodes would be contained 3 nodes to classify ozone concentrations into 3 groups of air quality: good, moderate and



unhealthy air. Hence, the network architecture to be investigated for this study was illustrated in Fig. 1.

Figure 1: A neural network predictive model for classifying of ozone concentrations

4.4. Comparison of Model Performance

To compare the performance of the discriminant analysis model with the neural network model, the CCR and the average CCR of predictive accuracies in discriminant analysis (DA) and neural network (NN) are shown in Table 3.

Table 3 clearly displayed in general both the discriminant analysis and the neural network models can produce good classification results with higher 78% of the average CCR. In addition, the neural network model demonstrates a superior ability to categorize the groups of air quality as seeing the higher average CCR than the discriminant analysis model.

5. Conclusion and Discussion

To gain more understandings and implications, the conclusion derived from this study can be summarized as follows: (1) The influential variables impacted the ozone concentration in the east of Thailand through the values of factor loadings in factor analysis to be; Oxide of Nitrogen, CO&PM$_{10}$, Hydro Carbon

Put into Group	True Group							
	Good Air		Moderate Air		Unhealthy Air		Average CCR	
	DA	NN	DA	NN	DA	NN	DA	NN
Training Data Set								
Good Air	997	1,112	82	112	0	1		
Moderate Air	151	68	217	231	2	15		
Unhealthy Air	32	0	45	1	16	2		
Total of O_3	1,180	1,180	344	344	18	18		
CCR	0.8449	0.9424	0.6308	0.6715	0.8888	0.1111	0.7977	0.8722
Validation Data Set								
Good Air	472	516	36	59	2	1		
Moderate Air	48	21	89	109	2	12		
Unhealthy Air	17	0	47	4	10	1		
Total of O_3	537	537	172	172	14	14		
CCR	0.8790	0.9609	0.5174	0.6337	0.7143	0.0714	0.7898	0.8658

Table 3: Classification results of discriminant analysis and neural network models

Group, NMHC, SO_2, Temperature and Sun Radiation, Pressure, Rain, Relative Humidity and Wind. All these variables are in agreement with the results obtained by Saithanu and Mekparyup [21]. (2) To achieve good generalization on unseen data, a validation data set is required other than a training data set. The validation data set provided more classification accuracy in classifying the group of good air for both of discriminant analysis and neural network models. (3) Neural network model is shown to perform well relative to discriminant analysis model in the prediction of the group of good air and the group of moderate air. However, discriminant analysis model is shown better performance than neural network model in categorizing the group of unhealthy air. That is, neural network method requires sufficient pairs of inputs and targets (outputs) for training the network.

Acknowledgments

We are grateful to National Research Council of Thailand (NRCT) for supporting this research grant and also to the Air Quality and Noise Management Bureau, Pollution Control Department for kindly providing all data.

References

[1] R. Vecchi, G. Valli, Ozone assessment in the southern part of the Alps, *Atmos Environ*, **33**, No. 1 (1999), 97-109.

[2] C. Duenas, M. C. Fernandez, S. Canete, J. Carretero, E. Liger, Assessment of ozone variations and meteorological effects in an urban area in the Mediterranean Coast, *The Science of the Total Environment*, **299** (2002), 97-113.

[3] URBANEMISSIONS, Info, Air Quality Index Calculator. Retrieved January, 13, 2013, from http://www.urbanemissions.info/model-tools/aqi-calculator.html

[4] Office of Natural Resources and Environmental Policy and Planning, Notice of the National Environment Committee NO.28 (B.E.2550) on Air Quality Standards. Retrieved October 25, 2011, from http://www.legalbase.pti.org/ Law.aspx?lid=3956

[5] P. Khaenamkaew, P. Iamraksa, S. Raksawong, K. Wongsontam, C. Angwanisakul, S. Khuntong, Annual Concentration Report and Emission Sources Analysis of the Air Pollutants Measured by the AQM Station, Kasetsart University, Si Racha, Thailand, *Research Exhibition "Research in Kasetsart University 2011" in National Agricultural Fair* (2011).

[6] Thai Universities for Healthy Public Policies, 2011, Air quality around Map Ta Phut in Rayong, Thailand. Retrieved November 13, 2012, from www.ehwm.chula.ac.th/maptaphut/air-maptaphut.pdf

[7] L. Malec, F. Skacel, T. Fousek, V. Tekac, P. Kral, Analyzing ground ozone formation regimes using a principal axis factoring method: A case study of Kladno (Czech Republic) industrial area, *Atmosfera*, **21**, No. 3 (2008), 249-263.

[8] A. Lengyel, K. Heberger, L. Paksy, O. Banhidi, R. Rajko, Prediction of ozone concentration in ambient air using multivariate methods, *Chemosphere*, **57** (2004), 889-896.

[9] C. Ghiaus, F. Caini, R. Belarbi, Linear discriminant analysis applied to forecast ozone concentration classes in sea-breeze regime, *Fifth International Conference on Urban Climate* (2003).

[10] M. W. Gardner, S. R. Dorling, Meteorologically adjusted trends in UK daily maximum surface ozone concentrations, *Atmospheric Environment*, **34** (2000), 171-176.

[11] S. A. Abdul-Wahab, S. M. Al-Alawi, Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks, *Environment Modelling & Software*, **17**, No. 3 (2002), 219-228.

[12] S. I. V. Sausa, F. G. Martins, M. C. M. Alvim-Ferraz, M. C. Pereira, Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations, *Environmental Modelling & Software*, **22**, No. 1 (2007), 97-103.

[13] S. Hassanzadeh, F. Hosseinibalam, M. Omidvari, Statistical methods and regression analysis of stratospheric ozone and meteorological variables in Isfahan, *Physica A*, **387**, No. 10 (2008), 2317-2327.

[14] L. H. Nghiem, N. T. Kim Oanh, Comparative analysis of maximum daily ozone levels in urban areas predicted by different statistical models, *ScienceAsia*, **35** (2009), 276-283.

[15] R. A. Johnson, D. W. Wichern, *Applied Multivariate Statistical Analysis, 6-th Edition*, Prentice-Hall Press, NJ, USA (2007).

[16] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Mathematics of Control, Signals, and Systems*, **2**, No. 4, 303-314.

[17] K. Hornik, M. Stinchcombe, H. White, Multilayer feed forward networks are universal approximators, *Neural Networks*, **2**, No. 5 (1989), 359-366.

[18] Y. Guo, K. J. Dooley, Identification of Change Structure in Statistical Process Control, *International Journal of Production Research*, **30** (1992), 1655-1669.

[19] E. A. El-Sebakhy, A. S. Hadi, K. A. Faisal, Iterative Least Squares Functional Networks Classifier, *IEEE Transactions on Neural Networks*, **18**, No. 3 (2007), 844- 850.

[20] C. Oh, S. G. Ritchie, Recognizing vehicle classification information from blade sensor signature, *Pattern Recognition Letters*, **28**, No. 9 (2007), 1041-1049.

[21] K. Saithanu, J. Mekparyup, Clustering of Air Quality and Meteorological Variables Associated with High Ground Ozone Concentration in the Industrial Areas, at the east of Thailand, *International Journal of Pure and Applied Mathematics*, **3**, No. 81 (2012), 505-515.