Modeling Haze Problems in the North of Thailand using Logistic Regression

Busayamas Pimpunchat[1], Khwansiri Sirimangkhala[1] & Suwannee Junyapoon[2]

[1]Industrial Mathematics Research Unit & Department of Mathematics,
Faculty of Science, King Mongkut's Institute of Technology Ladkrabang,
Bangkok 10520, Thailand
[2]Department of Chemistry, Faculty of Science,
King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand.
Email: knbusaya@kmitl.ac.th

Abstract. At present, air pollution is a major problem in the upper northern region of Thailand. Air pollutants have an effect on human health, the economy and the traveling industry. The severity of this problem clearly appears every year during the dry season, from February to April. In particular it becomes very serious in March, especially in Chiang Mai province where smoke haze is a major issue. This study looked into related data from 2005-2010 covering eight principal parameters: PM_{10} (particulate matter with a diameter smaller than 10 micrometer), CO (carbon monoxide), NO_2 (nitrogen dioxide), SO_2 (sulphur dioxide), RH (relative humidity), NO (nitrogen oxide), pressure, and rainfall. Overall haze problem occurrence was calculated from a logistic regression model. Its dependence on the eight parameters stated above was determined for design conditions using the correlation coefficients with PM_{10}. The proposed overall haze problem modeling can be used as a quantitative assessment criterion for supporting decision making to protect human health. This study proposed to predict haze problem occurrence in 2011. The agreement of the results from the mathematical model with actual measured PM_{10} concentration data from the Pollution Control Department was quite satisfactory.

Keywords: *forecasting; haze problem; multivariate logistic regression; mathematical model; PM_{10}.*

1 Introduction

For the past decades, a thick particle haze that covers the Northern region of Thailand, i.e. Chiang Mai, Chiang Rai, Lamphun, Lampang, Mae Hong Son, Phayao, Phrae and Nan, has been causing serious air pollution problems during the dry season. [1] A haze crisis caused by the dry and stagnant weather conditions occurs every year in the dry season, from February to April, especially in March. These conditions produce dust particles smaller than 10 microns (PM_{10}) suspended in the atmosphere. During this period, a large amount of particulate matters are released into the atmosphere, including carbon monoxide (CO), carbon dioxide (CO_2), volatile organic compounds (VOCs),

Received September 24th, 2013, Revised May 30th, 2014, Accepted for publication June 30th, 2014.
Copyright © 2014 Published by ITB Journal Publisher, ISSN: 2337-5760, DOI: 10.5614/j.math.fund.sci.2014.46.2.7

carcinogenic polycyclic aromatic hydrocarbons (PAHs) [2]. The main emission source is biomass open burning, such as forest fires, solid waste burning, and agricultural residue field burning [3,4]. The air pollutants are trapped near ground level due to the meteorological conditions (e.g. stagnant air) and the basin-liked topography surrounded by high mountain ranges results in restricted pollution dispersion. In addition, low rainfall in dry season effect on violence of haze problem. For this reason, the leaching of smoke or dust particles in the air is low [4]. A report from the Pollution Control Department revealed that Chiang Mai province encounters ever longer and more serious smog problems.

During haze periods, the 24-hour average concentrations of particulate matter with an aerodynamic diameter smaller than 10 micron (PM_{10}) frequently exceed the Thailand National Ambient Air Quality Standard (NAAQS) of 120 micron [5]. Also, the Public Health Ministry of Thailand [1] has reported an increase in bronchial asthma and respiratory diseases in people living in these areas. In addition, these fine particles contain carcinogenic polycyclic aromatic hydrocarbons that can induce lung cancer [6]. The smoke haze episodes affect not only adverse health problems but also reduce visibility and cause economic sectors to decline, i.e. tourism, transportation, and agriculture. Therefore, the Thai government shows increasing concern to get the smoke haze problem under control [5].

The PM_{10} measurement method is complex and it takes a long time to analyze the results with a performance tool [7]. In order to simplify this process, we utilized a logistic regression analysis method for establishing the occurrence of haze. As the first step, we determined the main factors associated with the occurrence of haze in the Chiang Mai area. The second step was to establish a mathematical model to represent the occurrence of haze using the main factors from the logistic regression analysis. Lastly, we used the mathematical model to predict the occurrence of haze in the next year. This study can be used as guidance for monitoring haze problems in the northern region of Thailand.

2 Study Data

This study is based on an investigation of hourly measured data over a six-year period, from 2005 to 2010. PM_{10} is one of five air pollutants that are used to measure air quality in Thailand. The Air Quality Index (AQI) is a number used by government agencies to communicate to the public how polluted the air is currently or how polluted it is forecasted to become.

2.1 Air Quality Index

The Air Quality Index (AQI) is an indicator of air quality, based on air pollutants that have adverse effects on human health and the environment. The main pollutants for Thailand are: O_3, SO_2, NO_2, CO and PM_{10}. The maximum AQI from these five pollutants is an indicator of the air quality for a particular day or moment. The Pollution Control Department of the Ministry of Natural Resources and Environment is the organization in Thailand that measures the AQI and reports the results [8,9].

The effects of air pollution on people's health can become a cause for concern, as shown Table 1. During the dry season of 2007, there was no rainfall for 6 months (November 2006 – April 2007). The occurrence of smoke haze affected a high PM_{10} concentration of up to 303.9 micron [4]. The number of patients infected with respiratory diseases and allergies increased dramatically according to a report from the Public Health Department of Chiang Mai province.

Table 1 Air Quality Index criteria in Thailand.

AQI Values	Level of Health Concern	Colors	Meaning
0-50	Good	Blue	Air quality is considered satisfactory, and air pollution poses little or no risk.
51-100	Moderate	Green	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
101-200	Unhealthy	Yellow	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
201-300	Very unhealthy	Orange	Everyone may begin to experience health effects: members of sensitive groups may experience more serious health effects.
Over 300	Hazardous	Red	Health alert: everyone may experience more serious health effects.

Source: Pollution Control Department (www.pcd.go.th)

2.2 PM_{10}

PM_{10} is suggested as an indicator with relevance to the majority of the epidemiological data. It is extensively measured throughout the world [10,11]. As mentioned earlier, knowing the occurrence of PM_{10} from the ambient parameters is significantly necessary. The standard level of PM_{10} in Thailand is declared at 120 micron. However, during haze occurrence, 100 micron is set to monitor and alert people. In this study, PM_{10} was investigated in order to categorize haze occurrence response into 3 levels (0, 1 and 2) using logistic regression analysis. The detection limits were 0-99, 100-119, and over 120 micron, respectively indicating: little health effect; more health effect and alert; more serious health effect. PM_{10} was implicated with associated factors for predicting the occurrence of haze.

3 Methodology

The procedure of this research was as follows:

Step 1: Collect data of PM_{10}, CO, NO, NO_2, SO_2, O_3, wind direction, temperature, relative humidity (RH), pressure, wind speed and rainfall.
Step 2: Find the relationships between PM_{10} and others parameters by Pearson correlations for relationship classification into high, moderate, or low.
Step 3: Construct logistic regression for haze problem model using high and moderate relationships. Define response function on the basis of PM_{10} concentrations, reflecting the different ranges as: "little health effect (0-99)", "more health effect and alert (100-119)" and "more serious health effect (over 120)".
Step 4: Analysis and discussion of logistic regression model. Present goodness of fit by Pseudo R^2.
Step 5: Consider level of health concern for PM_{10} levels in 2011 to indicate health effects.

3.1 Logistic Regression Analysis

Logistic regression or logit regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable based on one or more predictor variables. That is, it is used for estimating the empirical values of parameters in a qualitative response model. The probabilities describing the possible outcomes of a single trial are modeled as a function of the predictor variables using a logistic function. Usually, *logistic regression* is used to refer specifically to problems in which the dependent variable is binary, which means that the number of available categories is two. Problems with more than two categories are referred to as *multinomial logistic regressions* or, if the multiple categories are ordered, as *ordered logistic regressions* [12,13]. Moreover, logistic regression measures the relationship between a categorical dependent variable and one or more independent variables that are usually continuous. The probability scores are utilized as the predicted values of the dependent variable [14].

3.2 Mathematical Model

Haze smoke problems normally occur in Northern Thailand for a short period from February to April each year, peaking in March. The factors relating to this problem correspond to 11 parameters [9,12], i.e. PM_{10}, CO, NO, NO_2, SO_2, O_3, wind direction, temperature, relative humidity (RH), pressure, wind speed, and rainfall. In order to identify the relationships between PM_{10} and the others parameters, the Pearson correlations were analyzed.

3.2.1 Model Formulation

The objective of modeling the relation between the occurrence of haze (dependent variable) and concentrations of gases with meteorological data (independent variables) is to estimate the probability of the former given the incidence of the latter. An explanation of logistic regression begins with an explanation of the logistic response function, which always takes on values between zero and one. Let P_i denote the multinomial probability of an observation falling in the j th category. Category j is used as the base level of the method (any category can be taken as the base level). We want to find the relationship between this probability and n explanatory variables, $x_1, x_2, ..., x_n$. The multinomial logistic regression model then is

$$\ln\left(P_i / P_j\right) = \beta_{0i} + \beta_{1i}x_1 + ... + \beta_{ni}x_n \tag{1}$$

and $P_1 + P_2 + ... + P_j = 1$ (2)

where $\beta_0, \beta_1, ..., \beta_n$ are model parameters for $i = 1, 2, ..., m$.

Since all the P 's add to unity, this can be reduced to

$$P_i = \frac{\exp\left(\beta_{0i} + \beta_{1i}x_1 + ... + \beta_{ni}x_n\right)}{1 + \sum_{k=1}^{j-1} \beta_{0k} + \beta_{1k}x_1 + ... + \beta_{nk}x_n} \tag{3}$$

For $i = 1, 2, ..., j-1$

$$P_j = \frac{1}{1 + \sum_{k=1}^{j-1} \beta_{0k} + \beta_{1k}x_1 + ... + \beta_{nk}x_n} \tag{4}$$

The model parameters are estimated by the method of maximum likelihood [12,13].

3.2.2 Estimation of Model Parameters in Logistic Regression

The regression coefficients are usually estimated through an iterative maximum likelihood method. Unlike linear regression with normally distributed residuals, it is not possible to find a closed-form expression for the coefficient values that maximizes the likelihood function, so an iterative process must be used instead [13].

The log likelihood functions for the multinomial logistic regression model are:

$$\ln L = \sum_{i=1}^{N}\sum_{r=1}^{j-1}\left(y_{ir}\sum_{k=0}^{K}\beta_{kr}x_{ik} - n_i \ln\left(1 + \sum_{r=1}^{j-1}\exp\left(\sum_{k=0}^{K}\beta_{kr}x_{ik} \right)\right)\right) \tag{5}$$

where Y_{ir} is a known, fixed constant value. Logistic regression models require a minimum of log likelihood to find parameters $\beta_0, \beta_1, ..., \beta_n$. Since Eq. (5) is nonlinear, Newton's method is used for this model. [14]

3.2.3 Measurement of Relationships

In linear regression, the squared multiple correlation R^2 is used to assess goodness of fit, as it represents the proportion of variance in the criterion that is explained by the predictors. In logistic regression analysis, there is no agreed upon analogous measure but there are several competing measures, each with its own limitations [12]. Pseudo R^2 is used to assess goodness of fit, as it represents the proportion of variance in the criterion that is explained by the predictors in the logistic regression analysis. The Pseudo R^2 formula used here is:

$$Pseudo\ R^2 = \frac{-2\log L_{Null} - [2\log L_{Model}]}{-2\log L_{Null}} \tag{6}$$

where L_{Null} is a likelihood function with only constants

L_{Model} is a likelihood function with the predicted value defined

Pseudo R^2 always takes on values between zero and one.

4 Results and Discussion

PM$_{10}$ is an important variable that indicates the violence of a haze situation. In this study, we firstly found the correlation coefficient between PM$_{10}$ and the concentrations of gases along with meteorological factors to formulate a model for predicting the occurrence of haze using a logit function.

4.1 Results of Correlation Coefficients

Correlation analysis was used to approach the relationship between two or more variables by correlation coefficient (r) values from –1 to 1. A negative value represents a contrary relationship, whereas a positive value shows the same direction among the sets of variables [12]. The values were defined as follows: (i) r from 0.50 to 1.00 or from –0.50 to –1.00 was considered a high correlation; (ii) from 0.30 to 0.49 or –0.30 to –0.49 was considered a moderate correlation; (iii) from 0.10 to 0.29 or –0.10 to –0.29 was considered a low correlation; and (iv) zero represents no relationship.

From the above principle, we obtained seven main factors associated with the concentration of PM_{10}, derived from the high and moderate correlation values, as shown in Table 2. Thus, the main factors associated with the PM_{10} concentration are CO, NO_2, O_3, RH, SO_2, NO and pressure. We tried using only the high correlation group, which resulted in less accuracy than using both the high and moderate correlations.

Table 2 Correlation coefficients of all parameters.

Factors	Correlation	Factors	Correlation
CO	0.6676	Temperature	0.0160
NO	0.3191	RH	-0.6633
NO_2	0.8245	Pressure	0.3278
SO_2	0.4719	Wind Speed	-0.0642
O_3	0.5576	Wind Direction	-0.0865
Rainfall	-0.2234		

In contrast to the independent variables (X_i), which are quantitative variables, the dependent variable (Y) is a qualitative variable. Moreover, the values of Y were coded for the smog problems into 3 cases. The response function on the basis of PM_{10} concentrations reflects the different ranges as shown in Table 2 , where all the values of CO, NO_2, O_3, RH, SO_2, NO are in $\mu g/m^3$ and pressure is in bar.

Let P_1 be the probability that $0 \leq PM_{10} \leq 99$, or, "little health effect"; P_2 is the probability that $100 \leq PM_{10} \leq 119$, or, "more health effect and alert"; and P_3 is the probability that $PM_{10} \geq 120$, or, "more serious health effect". Category 1 is the base level in our description of the method.

The logistic regression models of this study are:

$$\ln\left(P_2 / P_1\right) = \beta_{02} + \beta_{12}x_1 + \ldots + \beta_{72}x_7 \quad (7)$$

$$\ln\left(P_3 / P_1\right) = \beta_{03} + \beta_{13}x_1 + \ldots + \beta_{73}x_7 \tag{8}$$

and $P_1 + P_2 + P_3 = 1$

This can be reduced to:

$$P_1 = \frac{1}{1 + \sum_{k=2}^{3} \beta_{0k} + \beta_{1k}x_1 + \ldots + \beta_{7k}x_7} \tag{9}$$

$$P_2 = \frac{\exp\left(\beta_{02} + \beta_{12}x_1 + \ldots + \beta_{72}x_7\right)}{1 + \sum_{k=2}^{3} \beta_{0k} + \beta_{1k}x_1 + \ldots + \beta_{7k}x_7} \tag{10}$$

and $P_3 = \dfrac{\exp\left(\beta_{03} + \beta_{13}x_1 + \ldots + \beta_{73}x_7\right)}{1 + \sum_{k=2}^{3} \beta_{0k} + \beta_{1k}x_1 + \ldots + \beta_{7k}x_7}$ (11)

where $x_1, x_2, \ldots, x_7$ are concentration of CO, NO_2, O_3, RH, SO_2, pressure and NO, respectively.

4.2 Results of Coefficient Estimation

The logistic regression models are

$$\begin{aligned}\ln\left(P_2 / P_1\right) = {} & 0.896x_1 + 0.205x_2 - 0.052x_3 + 0.068x_4 + 0.113x_5 \\ & -0.114x_6 + 0.256x_7 - 89.639\end{aligned} \tag{12}$$

$$\begin{aligned}\ln\left(P_3 / P_1\right) = {} & 2.232x_1 + 0.362x_2 - 0.103x_3 + 0.136x_4 + 0.242x_5 \\ & -0.136x_6 - 0.06x_7 - 188.344\end{aligned} \tag{13}$$

Considering the seven predictive variables that influence the occurrence of smog derived from PM_{10} concentration and considering the predictive ability of the logistic model, the variation of this haze problem model is 59.6% (Pseudo $R^2 = 0.5960$). The accuracy of the model was 94.66%.

Figure 1 shows the multinomial logistic regression of this model for 2011. The accuracy of the model for 2011 was 92.33%. Figure 2 shows when high concentrations of PM_{10} affecting the haze situation advise monitoring and forecasting of criteria pollutants in the air during March and April in 2011. From the smog haze prediction for 2011 it can be seen that it is an important task to make how bad or good the air quality is for human health easily understandable and to assist in data interpretation for the decision making processes related to pollution migration measures and air quality management.

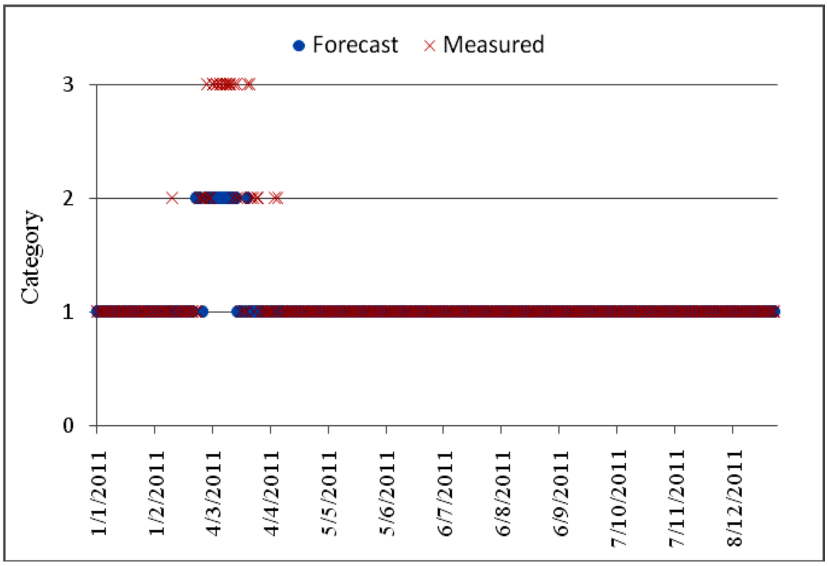

Figure 1 Multinomial logistic regression model in 2011.

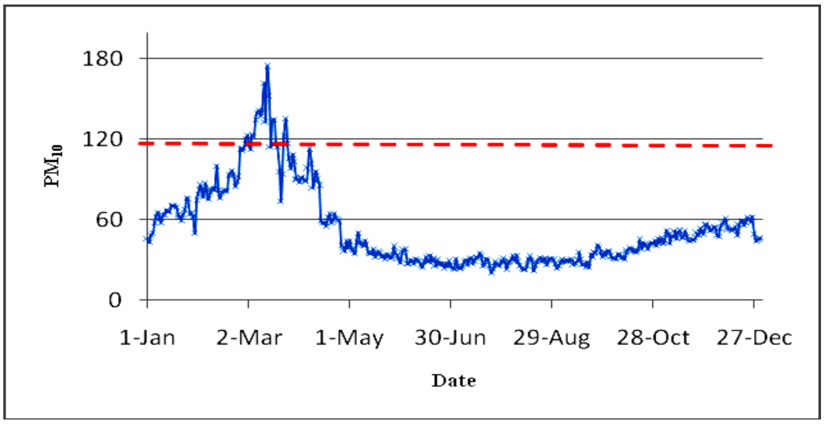

Figure 2 Plot of PM_{10} daily concentrations from PCD, Thailand in 2011.

Figure 3 shows that smog haze situations affecting the level of human health concerns could be classified into 3 levels of air quality: “Good” about 60% (219/365), “Moderate” about 31.23% (114/365), and “Unhealthy” about 8.77% (32/365). A comparison between the observed values and the model’s predicted values suggests that the model can be used for the prediction of daily PM_{10} concentrations in urban areas.

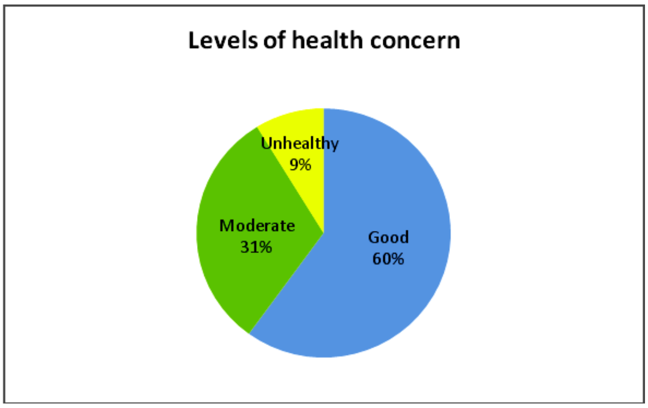

Figure 3 Level of Health Concern of PM_{10} in 2011.

5 Conclusions

A modeling effort was conducted in order to investigate human health concerns related to smoke haze for decision-making purposes. Logistic regression analysis was used as a tool for achieving the difficult task of predicting daily PM_{10} concentrations based on the main parameters affecting smog haze situations in Chiang Mai province. As a result it is believed that the derivation of PM_{10} concentrations from the parameters affecting smog haze situations should be considered as a tool for operational use in PM_{10} concentration forecasting, aiding the protection of the exposed population against short-term variations in particulate matter levels.

Acknowledgements

The authors would like to thank the Pollution Control Department in Thailand, which kindly provided valuable data. The financial support for this research project from the faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Thailand is gratefully acknowledged.

References

[1] PCD (Pollution Control Department), http://www.pcd.go.th (6 July 2012).

[2] Phoothiwut, S. & Junyapoon, S., *Size Distribution of Atmospheric Particulate-bound Polycyclic Aromatic Hydrocarbons and Characteristics of PAHs during Haze Period in Lampang Province, Northern Thailand*, Air Quality, Atmosphere & Health, **6**(2), pp. 397-405, 2013.

[3] Chantara, S., Sillapapiromsuk, S. & Wiriya, W., *Atmospheric Pollutants in Chiang Mai (Thailand) over a Five-year Period* (*2005-2009*)*, their Possible Sources and Relation to Air Mass Movement*, Atmospheric Environment, **60**, pp. 88-98, 2012.

[4] Wiriya, W., Sillapapiromsuk, S. & Chantara, S., *PM10-Bound Polycyclic Aromatic Hydrocarbons in Chiang Mai (Thailand): Seasonal Variations, Source Identification, Health Risk Assessment and Their Relationship to Air-Mass Movement*, Atmospheric Research, **124**, pp. 109-122, 2013.

[5] PCD (Pollution Control Department), Maryland Department of the Environment, & Chiang Mai Municipality, *Chiang Mai Emission Inventory in Municipality and Neighborhood Area*, Report of Ability of Federal and Local Government Official, Chiang Mai, 2002.

[6] Pengchai, P., Chantara, S., Sopajaree, K. & Wangkarn, S., Tengcharoenkul, U. & Rayanakorn, M., *Seasonal Variation, Risk Assessment and Source Estimation of PM_{10} and PM_{10}-bound PAHs in the Ambient Air of Chiang Mai and Lumphun, Thailand*, Environ. Monit. Assess., **154**, pp. 197-218, 2009.

[7] Pollution Control Department, *Manual Measurement of Dust in Ambient*, Bangkok: Kochakorn Publishing, 2003.

[8] Rayanakorn, M., *Haze and Air Pollution in Chiang Mai*, Chiang Mai: Login Design Work, pp. 9-14, 2010.

[9] Jeremy, C., *Air Pollution: An Introduction*, London: E&FN Spon, 1997.

[10] Kumer, A. & Goyal, P., *Forecasting of Daily Air Quality Index in Delhi*, Science of the Total Environment, **409**, pp. 5517-5523, 2011.

[11] World Health Organization, *WHO Air Quality Guidelines Global Update 2005*, Report on Working Group Meeting, Bonn, Germany, 18-20 October 2005.

[12] Kleinbaum, D.G., *Logistic Regression: A Self-Learning Text,* 3rd ed., New York: Springer, 2010.

[13] Hilbe, J.M., *Logistic Regression Models*, Boca Raton: CRC Press, 2009.

[14] Hamilton, L.C., *Statistics with Stata*, Pacific Grove, CA: Brooks/Cole Publishing Company, pp. 137-145, 1990.