Eco. Env. & Cons. 26 (1) : 2020; pp. (441-445)
Copyright@ EM International
ISSN 0971–765X

Estimating of precipitation occurrence during 2006-2016 in Bangkok, Thailand

Wandee Wanishsakpong[13], Rhysa McNeil[23*] and Boonorm Chomtee[1]

[1]Faculty of Science, Kasetsart University, 10900, Bangkok, Thailand
[2]Faculty of Science and Technology, Prince of Songkla University, Pattani, 94000, Thailand
[3]Center of Excellence on Mathematic, CHE, Si Ayuthaya RD., 10400, Bangkok, Thailand

(Received 10 August, 2019; accepted 9 October, 2019)

ABSTRACT

This study aims to investigate factors related to the average monthly precipitation in Bangkok, Thailand using a logistic regression model. The data used in this study were monthly averaged data from the Thai Meteorological Department over an eleven-year period (2006-2016). The dependent variable was precipitation which was divided into two groups, namely the precipitation event and the no-precipitation event. Four independent variables employed in the analysis were relative humidity, temperatures, wind direction and wind speed. A logistic regression model showed that factors affecting precipitation occurrence were relative humidity greater than 73% and wind direction from the Southwest. The area under the curve (AUC) indicated that classification accuracy of logistic regression model was satisfactory.

Keywords*: Logistic regression model, Sum contrasts, Classification, Confidence interval, Precipitation*

Introduction

Climate change is causing various impacts, especially the problems of global warming and change in precipitation levels. The quantity of precipitation is not going down or up the same in all places on the earth's surface (Radzuan *et al*., 2013). The event of precipitation occurrence is an important outcome to evaluate with outcome effects of natural disasters such as flooding and drought. The fluctuation of precipitation in Thailand has declined over the past 50 years but since 2000, precipitation has been extreme. At present, Thailand faces problems of precipitation events and suffered from extreme flooding in 2011 (Yoon *et al*., 2015). Bangkok, Thailand's capital located in the central part of the country was especially affected by a 2011 flooding event (Arifwidodo and Tanaka 2015). Average annual precipitation in Bangkok is 1450.3 millimeters. On average, the month with the most precipitation is September (Department of Drainage and Sewerage, 2011). In 2011, the precipitation increased dramatically in May with this event causing various impacts on agriculture, health and transportation.

Many scientists have developed mathematical models for atmospheric dynamics over the years and have studied models in different situations in order to explain these significant changes (Yoon *et al*., 2015). Moreover, there are various techniques in machine learning such as neural networks, random forests, decision tree, logistic regression and others (Radzuan *et al*., 2013) to improve over standard linear regression for precipitation prediction (Applequist, 2002).

The objectives of this study were to estimate percentages of precipitation in Bangkok by using sum contrasts for logistic regression model. This model showed confidence intervals for comparing means

that did not involve selecting a reference group. In addition, factors related to the precipitation in Bangkok were studied, and the efficiency of classification was explained by using the receiver operating characteristic curves (ROC).

Methodology

Data

Monthly averaged data of precipitation, humidity, temperature, wind direction and wind speed of Bangkok metropolis station were obtained from the Thai Meteorological Department (Thai Meteorological Department). The considered data set ranged from January 2006 through December 2016 (132 months). The dependent variable used in this study was precipitation which was divided into two groups (precipitation event/no -precipitation event). Precipitation events were identified based on Thailand's rainfall (daily) criteria (Thai Meteorological Department). A precipitation event was defined by average rainfall greater than at least 3 millimeters, and a no-precipitation event was defined by average rainfall less than 3 millimeters. The four determinant factors were temperature, humidity, wind direction and wind speed which related to the level of precipitation. Each of the four determinant factors was classified into two levels based on overall average of these factors. The temperature groups were less than or equal to 29 °C and more than 29°C. The relative humidity groups were less than or equal to 73% and more than 73%. The wind direction groups were Southeast winds and Southwest winds. The wind speed groups were less than or equal to 10 knots (Light/gentle breeze) and more than 10 knots (Moderate breeze) (Thai Meteorological Department).

Statistical methods

The logistic regression model is a method that has fewer assumptions than the linear discriminant model, the results can be given clearly, and the regression coefficient can be easily defined (Kiang, 2003; Radzuan *et al.*, 2003). This model formulated the logit of the probability *p* that a precipitation event as an additive linear function of the four determinant factors as follows:

$$\log\left(\frac{P}{1-P}\right) = \mu + \alpha_i + \beta_j + \gamma_k + \tau_k \quad .. (1)$$

where μ is constant α_i, β_j, γ_k and τ_k refer to temperature group, humidity group, wind direction group and wind speed group (Hosmer and Lemeshow, 2000; Kleinbaum and Klein, 2002).

The model (1) provides confidence interval for percentage of precipitation event for levels of each determinant. The confidence intervals were adjusted for other determinants using sum contrast method (Tongkumchum and McNeil, 2009; Kongchovy and Sampantarak, 2010).

A sum contrast was used to obtain confidence intervals for comparing mean/proportions with overall mean/proportions. The confidence intervals provided a criterion for classifying levels of the factor into three groups according to whether the confidence interval exceeded, crossed or was below the overall mean (Pipatjaturon *et al.*, 2016). Confidence intervals based on sum contrasts are more appropriate than the confidence intervals based on the treatment contrasts because commonly, the confidence interval based on treatment contrasts measures the difference from a reference group. Confidence intervals based on sum contrasts can be applied equitably to each category and compared percentage of precipitation event in each category factor with the overall percentage (Venables and Ripley, 2002; Tongkumchum and McNeil, 2009).

The Hosmer-Lemeshow test is a statistical test for goodness of fit for logistic regression. The Hosmer-Lemeshow test statistic is given by

$$\sum_{g=1}^{G} \frac{(O_{1g} - E_{1g})^2}{N_g \pi_g (1-\pi_g)} \quad .. (2)$$

where O_{1g}, E_{1g}, N_g, π_g and G denote the observed Y=1 events, expected Y=1 events, total observation, predicted risk for the g^{th} risk decide group and the number of groups, respectively. The test statistic asymptotically follows a distribution with G-2 degrees of freedom (Alan, 2012).

To assess the goodness of fit of the model the receiver operating characteristic (ROC) curve was used. ROC is comprised of graphical plots that represent the diagnostic ability of a binary classification (Pipatjaturon *et al.*, 2016). The ROC curve was created by plotting the true positive rate in the y-axis (i.e. predicting precipitation when it did precipitate) against the false positive rate in the x-axis (i.e. predicting precipitation when it did not precipitate).

ROC analysis provides tools to select possibly optimal models (Zweig *et al*., 1993) and show the trade-off in a model between correctly predicting precipitation when it truly does precipitate and predicting precipitation when it truly does not precipitate.

Denoting the predicted outcome as 1 (precipitation) if $p \geq c$, or o (no precipitation) if $p < c$, its plotted sensitivity is the proportion of positive outcomes correctly predicted by the model against the false positive rate (proportion of all outcome incorrectly predicted). In this case, c value was 0.51. Sensitivity and specificity of the model provide a cut-off point in the curve where the predicted precipitation agrees with the observed value in the data. The area under the ROC curve (AUC) is a summary statistic of demonstrable performance. AUC would be 1 to perfect prediction, non-predictive (AUC 0.5), less predictive ($0.5 < AUC < 0.7$), moderately predictive ($0.7 < AUC < 0.9$) and highly predictive ($0.9 < AUC < 1$) (Swets, 1988; Greiner *et al*., 2000; Vanagas, 2004).

Statistical modelling and graphical presentation were carried out by using R statistical software (R Development Core Team, 2018).

Results and Discussion

Preliminary Results

Monthly averaged data of precipitation, temperature, relative humidity, wind direction and wind speed were studied for 11 years. The overall average precipitation levels in each group of temperature, relative humidity, wind direction and wind speed are explained in Table 1.

Table 1 shows the overall average precipitation level for temperatures less than or equal to 29 °C and more than 29 °C were 6.02 mm. and 7.98 mm., respectively. The overall average precipitation level of relative humidity less than or equal to 73% and for more than 73% were 2.66 mm. and 10.48 mm., respectively. The overall average precipitation level by wind direction from the Southeast was 4.67 mm., while overall average precipitation level from Southwest winds was 9.04 mm. The overall average precipitation level for wind speeds for light/gentle breeze and moderate breeze were 7.68 mm. and 6.31 mm., respectively.

The percentage of assessed precipitation events in two temperature groups, two humidity groups, two wind direction groups and two wind speed groups are shown in Figure 1

Figure 1 shows a bar chart of percentage of precipitation events by temperature group, humidity group, wind direction group and wind speed group superimposed with adjusted percentage and their 95% confidence intervals from the model (1) based on sum contrasts. The horizontal line is average percentage 63.63%. The model suggests that the confidence intervals of percentage of precipitation event in humidity groups and two wind direction groups were different and statistically significant (P-value <

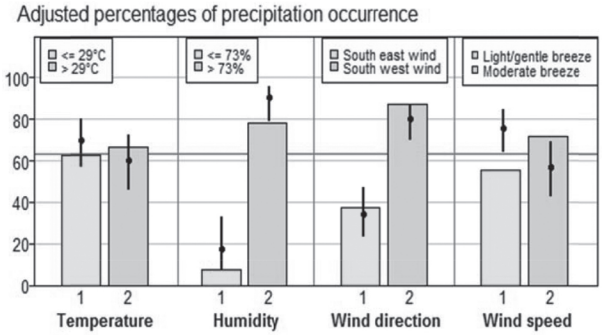

Fig. 1. The confidence interval of precipitation event by temperature, humidity, wind direction and wind speed

Table 1. Overall average precipitation level (mm) of temperature, relative humidity, wind direction and wind speed (2006-2016)

Factors	Categories	Overall Average Precipitation (mm) [SD]
Temperature	<= 29°C	6.02 [5.55]
	>29°C	7.98 [7.68]
Relative humidity	<=73%	2.66 [3.94]
	>73%	10.48 [6.53]
Wind direction	Southeast wind	4.67 [6.89]
	Southwest wind	9.04 [5.92]
Wind speed	Light /gentle breeze	7.68 [8.54]
	Moderate breeze	6.31 [4.26]

0.05). The confidence interval of percentage of humidity more than 73% and wind direction in Southwest wind exceeds the overall mean and the confidence interval of percentage of humidity in less than 73% and wind direction in Southeast wind below the overall mean.

Goodness of fit the logistic model

The chi-square test of the Hosmer-Lemeshow goodness of fit test was 5.131, which is statistically insignificant. It can be concluded that the model is appropriate. In addition, the ROC curve shows how well the model predicts a binary outcome as shown in Figure 2.

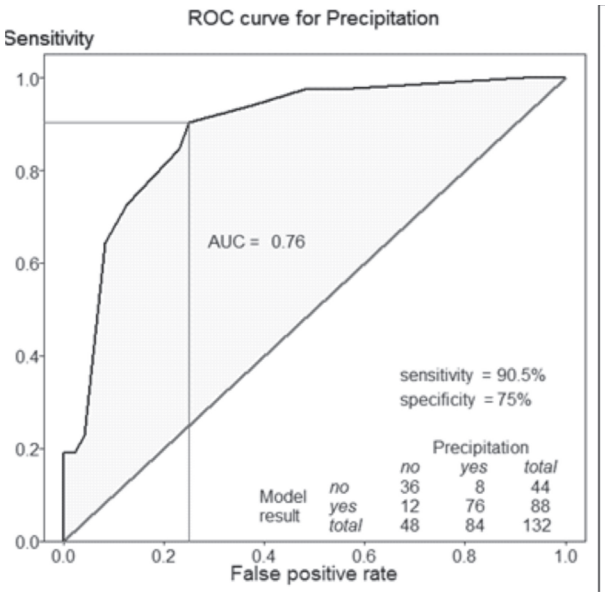

Fig. 2. Receiver Operating Characteristic (ROC) curve and cross classification observed with estimated outcome

Figure 2 shows the ROC curve of the logistic regression model. Choosing c=0.51 gives 84 predicted precipitation events, in agreement with the observed data. The red lines, drawn from the cut-off point to the x-axis and y-axis, show the model sensitivity and specificity (1-fase positive rate). The full model gives 90.5 % sensitivity, 75% specificity and AUC = 0.76. Therefore, the performance of the model was moderately predictive (Vanagas, 2004; Shen and Tan, 2005; Waeto *et al*., 2014).

The advantage of logistic regression is the technique chooses the significant variable. The result can be given clearly and regression coefficient can be defined easily (Kiang, 2003; Radzuan *et al*., 2013). Moreover, the logistic regression model provides the best relative performance. Although this advantage facilitates process forecast, it also has limitations. The method studied may be adversely affected when the underlying phenomenon is non-statistical (Kiang, 2003). However, there are many classification methods to employ for future study. Robust classification methods should be compared, such as neural network, decision tree learning, discriminant analysis and random decision forest. These method could help meteorologists select model variables to produce better precipitation forecasts (Radzuan *et al*., 2013).

Conclusion

An analysis of classification of precipitation levels. The dependent variable was conducted by dividing precipitation into two groups which were precipitation and no-precipitation. Four determinant factors were composed of relative humidity, temperature, wind direction and wind speed. The percentage of precipitation events in relative humidity more than 73% , wind direction in Southwest wind and wind speed more than 10 knots (moderate breeze) were higher than the average percentage 63.63%. The logistic regression model was used to study the factors affecting the precipitation, and the efficiency of classification was explained by the area under the ROC curve. The results of the logistic regression analysis showed that the relative humidity and wind direction effected precipitation. The area under curve (AUC) model showed a value close to 1. Therefore, the logistic regression model was accurate.

Acknowledgement

This research is support by the Center of Excellence in Mathematics, Commission on Higher Education,Thailand.

References

Alan, A. 2012. Categorical Data Analysis. Hoboken: Johy Wiley and sons.

Applequist, S., Gahrs, G.E., Peffer, R. and Niu, X. 2002. Comparison of

Methodologies for Probabilistic Quantitative Precipitation Forecasting. Weather and Forecasting. 17: 783-799.

Arifwidodo, S. and Tanaka, T. 2015. The Characteristics of

Urban Heat Islands in Thailand. Social and Behavioral Science. 195: 423-428. DOI : http://dx.doi.org/10.1016/j.sbspro.2015.06.484

Department of Drainage and Sewerage. 2011. Available at http://www.dds.bangkok.go.th

Greiner, M., Pfeiffer, D. and Smith, R. D. 2000. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Preventive Veterinary Medicine. 45(1-2): 23-41. DOI : http://dx.doi.org/10.1016/s0167-5877(00)00115X

Hosmer, D.W. and Lemeshow, S. 2000. Applied Logistic Regression. 2nd ed. New York: Springer-Verlag.

Kiang, M. Y. 2003. A Comparative assessment of classification methods. Decision Support Systems. 35: 441-454. DOI : http://dx.doi.org/10.1016/s0167-9236(02)00110-0

Kleinbaum, D.G. and Klein, M. 2002. Logistic Regression: A Self-Learning Text. 2nd ed. New York: Springer-Verlag.

Kongchovy, N. and Sampantarak, U. 2010. Confidence interval for adjusted proportion using logistic regression. Modern Applied Science. 4(6): 2-6.

Pipatjaturon, N., Ma-a-Lee, A., Thongkumchum, P. and Ueranantasan, A. 2016. Estimating Liver Cancer Death in Thailand : Methodologies to optimize the use of verbal Autopsy Data. Far East Journal of Mathematical Sciences. 10(100): 1595-1610. DOI : http://dx.doi.org/10.17654/MS100101595

R Development Core Team. 2018. R: A language and Environment for statistical Computing. R Foundation for statistical Computing: Vienna.

Radzuan, N. F. M., Putra, A., Othman, Z., Bakar, A. A. and Hamdan, A. R. 2013. Comparative Study-Three Artificial Intelligence Techniques for Rain Domain in precipitation Forest. *International Journal of Computer and Information Engineering.* 7(12): 930-935.

Shen, L. and Tan, E. C. 2005. Dimension reduction-based penalized logistic regression for cancer classification using microarray data. IEEE/ACM transactions on computational biology and bioinformatics/IEEE, ACM. 2(2): 166-175. DOI : http://dx.doi.org/10.1109/TCBB.2005.22

Swets, J. A. 1988. Measuring the accuracy of diagnostic systems Science. 240(4857): 1285-1293. DOI : http://dx.doi.org/10.1126/science.3287615

Tongkumchum, P. and McNeil, D. 2009. Confidence intervals using contrasts for regression model. Songklanakarin Journal Science Technology. 31(2): 151-156.

Vanagas, G. 2004. Receiver operating characteristic curves and comparison of cardiac surgery risk stratification systems. Interactive Cardiovascular and Thoracic surgery. 3(2): 319-322. DOI : http://dx.doi.org/10.1016/j.icvts.2004.01.008

Venables, W.N. and Ripley, B.D. 2002. Modern Applies Statistic with S. New York: Springer-Verlag.

Waeto, S., Pipatjaturon, N., Tongkumchum, P., Choonpradub, C., Saelim, R. and Makaje, N. 2014. Estimating Liver Cancer Deaths in Thailand based on Verbal Autopsy study. Journal of Research on Heath Science. 14(1): 18-22.

Yoon, S., Kumphon, B. and Park, J. 2015. Spatial modeling of extreme rainfall in northeast Thailand. Journal of Applied Statistics. DOI : http://dx.doi.org/10.1080/02664763.2015.1010492.

Zweig, M. H., Petrovich, G. N. and Prijovich, Z. M. 1993. ROC plots display test accuracy but are still limited by the study design. Clinical Chemistry. 39(6): 1345-1346.