Applications of Statistical Methods for Rainfall Prediction over the Eastern Thailand

Lily Ingsrisawang, Supawadee Ingsriswang, Pramote Luenam, Premjai Trisaranuwatana, Song Klinpratoom, Prasert Aungsuratana, and Warawut Khantiyanan

Abstract**—** This paper presents the use of three statistical methods: First-order Markov Chain, Logistic model, and Generalized Estimating Equation (GEE) in modeling the rainfall prediction over the eastern part of Thailand. Two daily datasets during 2004-2008, so-called Meteor and GPCM, were obtained from Thai Meteorological Department (TMD) and Bureau of the Royal Rain Making and Agricultural Aviation (BRRAA). The Meteor observation consists of the average of rain volumes (AVG) from 15 local weather stations, and the observation of the Great Plain Cumulus Model (GPCM) includes 52 variables, for example, temperature, humidity, pressure, wind, atmospheric stability, seeding potential, rain making operation, and rain occurrence. Merging and matching between the GPCM dataset and Meteor observations, the GPCM+Meteor dataset was generated including 667 records with 66 variables. The first-order Markov chain model was then built using the Meteor dataset to predict two transitional probabilities of a day being wet given the previous day being wet or being dry, P(W/W) and P(W/D), respectively. The odds ratio(OR) was computed from these probabilities and gave the value of 6.85, which indicated that it was about 7 times more likely to be a wet day given the previous day was also wet within the eastern region of Thailand, than that given the previous day was dry. Next, the logistic models were also fitted using the Meteor dataset by taking account of cyclical effect in modeling for the prediction of P(W/W) and P(W/D), respectively. The models showed that the odds ratios of being wet days are not constant over day *t* during the years 2004-2008. Finally, the GEE method was applied with the GPCM+Meteor dataset to study the effects of weather conditions on the prediction of rainfall estimates on wet days, by taking account of correlation structure among observations. The variables of -15 ^{o}c isotherm height and K-Index were shown statistically significant for the prediction of rainfall estimates at a 0.05 level. In order to effectively detect the rain conditions and make the right decisions in cloud-seeding operations, the statistical methods presented in this study can help in deriving the useful features from the rain and weather observations and modeling the rain occurrence.

Index Terms**— first-order Markov chain, generalized estimating equation, logistic model, seeding operation.**

I. Introduction

The agricultural areas in the eastern part of Thailand frequently face with severe drought every once a three-year period. It is also found that a lot of industries have been taken place and produced high pollution over the area. As a result, there is not enough water for residents and agriculturists in living, consumption, and agricultures [9]. It is necessary to enhance the precipitation in this area by conducting a number of cloud seeding operations under the royal rain making practical plan. However, there is no assurance for the success of cloud seeding operations, it is important to determine or forecast the success rate before any operations are conducted. Several climate factors, precipitation records, and prediction results from the cloud models such as the Great Plains Cumulus Model (GPCM) are normally used in making the decision on whether the cloud seeding operation will be launched or not [13]. Therefore, rainfall occurrence and rainfall estimates are our targets to evaluate the success of cloud seeding programs.

Several techniques both numerical modeling and machine learning have been studied for prediction of rainfall estimates using both radar and ground observations [1]-[3], [7], [8], [13]. In this paper, the methodologies of statistical methods such as First-order Markov Chain, Logistic model, and Generalized Estimating Equation (GEE) were employed to model the rainfall occurrence and rainfall estimates. The objectives of this study are to 1) apply various statistical methods in modeling the prediction of rainfall occurrence and rainfall estimates, and 2) identify what weather conditions affect the average amount of rainfall on wet days.

II. Materials and Methods

A. Data Preparation

Data that were required in this study consisted of 1) the upper air observations of 52 variables, including, temperature, humidity, pressure, wind, atmospheric stability, seeding potential, rain making operation, rain occurrence, and etc, which were derived from the GPCM, of Bureau of the Royal Rain Making and Agricultural Aviation, Thailand, and 2) the daily records of average of rain volumes (AVG) from 15 local weather stations which were obtained from Thai Meteorological Department. Two datasets, called Meteor and GPCM, were collected for the period of 2004 to 2008. The GPCM and the Meteor datasets for the eastern region contain 691 and 1,735 daily records, respectively. Based on the AVG variable, each record of the Meteor dataset was then categorized into rain or no-rain event by the following conditions: if any weather station had non-zero rain reported on day *t*, then the record on that day

Manuscript received January 19, 2010.

L. Ingsrisawang is with Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand (corresponding author: phone number: +66 2 5625555 EXT 4514; fax: +66 2 9428384; email: fscilli@ku.ac.th).

P. Trisaranuwatana is with Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand (email: fsciprt@ku.ac.th).

S. Ingsriswang is with National Center for Genetic Engineering and Biotechnology, Pathumthani, Thailand (email: supawadee@biotec.or.th).

P. Luenam is with School of Applied Statistics, National Institute of Development Administration, Bangkok, Thailand (email: pramote.l@ics.nida.ac.th).

S. Klinpratoom, P. Aungsuratana, and W. Khantiyanan are with [7]Bureau of the Royal Rainmaking and Agriculture Aviation, Bangkok, Thailand (e-mail: songkpt@hotmail.com; aungsuratana@yahoo.com; and warawutku@yahoo.com).

would be classified as rain. In opposite, if all weather stations had zero rain reported on day *t*, then the record on that day would be classified as no-rain. Moreover, the other interested outcome is the average of daily rain volume of wet days (AVGWD). The AVGWD was computed by taking the summation of rain volume on day *t* that was reported from each weather station and divided by number of weather stations which reported the non-zero amount of rainfall on day *t*. The GPCM+Meteor dataset was made by linking the GPCM dataset and the Meteor observations, consists of 667 daily records with total of 66 variables.

B. *Statistical Modeling*

Three statistical methods were employed in modeling the rainfall prediction using SAS 9.13, including the first-order Markov chain, the logistic regression model, and the GEE. The first two methods were suggested by Coe and Stern [2,3] and we broadly adopted for implementation with our data to predict the rainfall occurrence. For the last method, we applied its technique for modeling the relationship between the average of daily rain volume on wet days and the weather condition predictors. Therefore, the details of the study are as follows:

Firstly, we started with the simplest model for studying the pattern of occurrence of wet and dry days. The Meteor dataset was used to build the first-order Markov model for predicting two transitional probabilities of a day being wet: P(W/W) and P(W/D). The P(W/W) is the transitional probability of a day being wet, given that the previous day was also wet, while the P(W/D) is the transitional probability of a day being wet, given that the previous day was dry. Based on the transitional probabilities, the estimated odds ratios (OR) were calculated for indicating the chance of being two consecutively wet days over the eastern region of Thailand. According to the suggestion of Coe and Stern [2,3], the five data years on day t are treated as replicated observations. Each observation on day *t* is classified by dry or wet day and its previous day was also classified by dry or wet. The OR statistic was used to measure the association between the events of rainfall occurrence on day *t-1* and day *t*, which can be computed by the formula [11]:

$$\mathrm{OR} = \frac{\mathrm{P(W/W)}\text{莨}1-\mathrm{P(W/D)}]}{\mathrm{P(W/D)}\text{莨}1-\mathrm{P(W/W)}]}$$

Secondly, the logistic models were also fitted using the Meteor dataset by taking account of cyclical effect in modeling for the prediction of the P(W/W) and the P(W/D), respectively. The first harmonics function in terms of sin(2 π *t*/K) and cos(2 π *t*/K) were treated as explanatory variables for cyclical trend of each day *t* over the years. The logistic model is given by:

$$\log_e \left[\frac{\mathrm{P}_t \mid \text{rain event on day } t\text{-}1}{1\text{-}\ \mathrm{P}_t \mid \text{rain event on day } t\text{-}1} \right]$$
$$= \alpha_0 + \beta_{01} \sin(2\pi t / \mathrm{K}) + \beta_{02} \cos(2\pi t / \mathrm{K})$$

where P_t is the probability of wet day on day *t* given the previous day rain event,
and *t* is day of a year that can be 1,2,3…,K ; K=365 or 366 for leap year.

The model coefficients are estimated by the method of maximum likelihood and the significant test for individual coefficient is assessed by the Wald's chi-square statistics. In addition, the two logistic regression lines will be tested for the equality of the coefficients of harmonic terms by using the analysis of deviance [4], [11]. The difference of model deviances has a χ^2 - distribution with degrees of freedom equal to the difference in number of degree of freedoms [4]. If the chi-square statistics show statistically significant difference in the coefficients of the harmonic terms between the two regression lines, it will be indicated that the odds ratios of being wet days are constant over day *t*.

Thirdly, the GEE approach has been applied on the GPCM+Meteor dataset to model the average of daily rain volume on wet days (AVGWD). The GEE methodology was proposed by Liang and Zeger [10] for analyzing correlated data from longitudinal or panel data. For this study, the data set consists of 366 days (panels) and for each day *t* = 1, 2, 3,…, 366 there can be only 2, 3, 4, or 5 repeated observations of data years available. If we treat day *t* as day of a year, each day-panel is independent to each other. There are 366 independent day-panels and the repeated observations within day-panels are correlated. In GEE, it allows specifying the pattern of correlation structure among repeated observations within day-panel in terms of working correlation matrix such as independent, unstructured, exchangeable, or autoregressive [6]. The objective of GEE is to model the effects of covariates in the population on the continuous or discrete outcome variable accounting for the within-panel correlation through the use of marginal model. The marginal distribution of a response variable is assumed to follow a generalized linear model (GLM) in which the variance is the function of the mean [5], [10], [12]. The relationship between the marginal mean of the response, $\mathrm{E}(\mathrm{Y}_{ij}) = \mu_{ij}$, and explanatory variables, x_{ij} , is described by a known link function, g , in which $g(\mu_{ij}) = x'_{ij}\beta$. For example, the link functions for the binary and gamma response variables are defined as:

Logit link: $g(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right)$ for binary responses

Log link: $g(\mu_{ij}) = \log(\mu_{ij})$ for gamma response.

Therefore, the response variable AVGWD that is observed as continuous, positive, and skewed to the right is reasonable to be fitted as a gamma distribution with log link. Also, the first-order autoregressive, AR(1), correlation structure among repeated observations within day-panel is considered to be taken into account for modeling. The GEE model parameters are estimated from a marginal model by the method of quasi-likelihood estimation. The advantage of the GEE method is that its estimator is robust against misspecification of the working correlation matrix. The GEE estimation can be performed using PROC GENMOD with a REPEATED option in the SAS package.

III. Study Results

The first-order Markov model showed that the probability of a day being wet given the previous day was wet, P(W/W), in the eastern region was around 64%. In addition, the P(W/W) was estimated by season as following: 50%, 70%, and 43% for summer, rainy, and winter seasons, respectively. The probability of a day being wet given the previous day was dry, P(W/D), in the eastern region was 20% and the probabilities for summer, rainy, and winter seasons within the region were about 18%, 41%, and 7%, respectively. The OR was computed from these probabilities and gave the value of 6.85 at a 0.05 level of significance, which indicated that it was about 7 times more likely to be a wet day given the previous day was also wet within the eastern region of Thailand, than that given the previous day was dry.

The logistic models that were developed for the prediction of the P(W/W) and the P(W/D) by adding the first harmonics terms of sin(2 π t/K) and cos(2 π t/K) as explanatory variables for cyclical effects were shown in Table 1.

The two fitted logistic models were:

$$\log_e\left[\frac{\text{P(W/W)}}{\text{1- P(W/W)}}\right] = 0.2065 - 0.4245\sin(2\pi t / \text{K}) - 0.7472\cos(2\pi t / \text{K})$$

,

and

$$\log_e\left[\frac{\text{P(W/D)}}{\text{1- P(W/D)}}\right] = -1.3392 - 0.3184\sin(2\pi t / \text{K}) - 1.2593\cos(2\pi t / \text{K})$$

.

The coefficients of harmonic parameters were tested by the Wald chi-square statistics and turned out to be statistically significant at a 0.05 level with p-value < 0.0001 for prediction of the P(W/W) and the P(W/D), respectively. Additionally, the results of chi-square statistics indicated that the set of corresponding harmonic coefficients in both models were statistically significant difference at a 0.05 level with p-value < 0.0001. Next, the estimated ORs for each harmonic term in each model were computed by exp(β), and found that the ORs of being wet days were not constant over day t in the period of 2004 to 2008.

The results of fitting the univariate GEE model for prediction the AVGWD showed that the variables that were statistically significant at a 0.05 level consisted of the -15 °c isotherm height, K-index, sweat index, the relative humidity at convective condensation level, the average wind direction at the altitude of 1,000 -5,000, 5,000 -10,000, 10,000-15,000 feet, the average wind speed at the altitude of 1,000-5,000 feet, the average relative humidity at the altitude of 1,000 -5,000, 5,000 -10,000, and 20,000-25,000 feet. For the multivariate model, all the variables that had a p-value smaller than 0.20 were included with two more categorical variables of royal rain operation (yes/no) and warm cloud potential (poor/moderate/good). Based on a significance level 0.05, the significant predictors in the rainfall estimates model were the -15 °c isotherm height and K-index as presented in Table 2. Therefore, the prediction model of the AVGWD for the eastern region can be formulated as:

$$\begin{aligned}\log_e(\mu) = {} & -5.2819 - 0.1312(\text{royal rain operation}) \\ & + 0.0003(\text{-15}^{\circ}c \text{ isotherm height}) \\ & +0.0456(\text{K index}) - 0.0005(\text{sweat index}) \\ & - 0.0007(\text{the average wind direction} \\ & \text{at the altitude of 20,000-25,000 feet})\end{aligned}$$

where μ represents for the average of daily rain amount on wet days.

That is, the average of daily rain amount was estimated to increase by a factor of ($e^{100(0.0003)}$ -1) = 0.0304 or 3.04% as likely during the period 2004-2008 for every one- hundred unit increase in the -15 °c isotherm height while other variables in the model were held constant. Similarly, the estimate of average daily rain amount was likely to increase ($e^{(0.0456)}$ -1) = 0.0466 or 4.66% for every one unit increase in the K index while other variables were held constant.

IV. Conclusions

This study illustrated how we developed rainfall prediction models, starting from the simple Markov-chain model using only the data from amount of daily rainfall, moving to the Logistic model using the data from daily amount of rainfall and adding cyclical terms as predictors in model, and ending up with the GEE model using taking correlation structure into account for modeling the relationship of weather conditions on the prediction of rainfall estimates on wet days.

The GEE method is appropriate for analyzing this panel dataset of rainfall observations over the eastern Thailand. Not only we can apply the method for modeling the effects of weather condition covariates on the prediction of rain estimate, but also the prediction of rain occurrence by using different response distribution and different link functions. In SAS, there are choices of correlation structures that allow specifying for the repeated observations within day-panel. The solution of the GEE has the robustness property for statistically consistent parameter estimates and standard errors even if the correlation structure is misspecified or the data is missing completely at random (MCAR). Results from the GEE model indicated that the significant weather conditions in the averaged of daily rain volume on wet days for a period of 2004-2008 over the eastern part of Thailand were the -15 °c isotherm height and K-index. In summary, the statistical methods presented in this study can help in deriving the useful features from the weather observations and modeling the rain occurrence in order to effectively detect the rain conditions and make the right decisions in cloud-seeding operations.

Acknowledgment

The authors thank the Bureau of Royal Rainmaking and Agricultural Aviation (BRRAA) and the Thai Meteorological Department (TMD) as well as many individuals including the director and the administrators of the BRRAA and the director and the officers of the TMD for sharing information and supporting data for this study.

REFERENCES

[1] R.E. Chander, and H.S. Wheater, “Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland,” *Water Resour. Res., 38*(*10*), 2002, pp. 10-1– 10-11.

[2] R. Coe, and R.D. Stern, “Fitting Models to Daily Rainfall,.” *J. Appl. Meteorol., 21*, 1982, pp.1024-1031.

[3] R. Coe, and R.D. Stern, “A model fitting analysis of Daily Rainfall Data.”_*J.R. Statist. Soc. A, 147*, 1984. pp. 1-34.

[4] D. Collett, *Modelling binary data*. Boca Raton: Chapman & Hall/CRC, 1999.

[5] P. Diggle, P. Heagerty, K.Y. Liang, and S.L. Zeger, *Analysis of Longitudinal Data* (2nd ed.). Oxford: Oxford University Press, 2002.

[6] G.M. Fitzmaurice, N.M. Laird, and J.H. Ware, *Applied Longitudinal Analysis.* New Jersey: John Wiley & Sons, Inc., 2004.

[7] K.P. Gabriel, and J. Neumann, “A markov chain model for daily rainfall occurrence at Tel Aviv.,” *Quart. J. Roy. Metror. Soc., 88*, 1962, pp. 90-95.

[8] L. Ingsrisawang, S, Ingsriswang, S. Somchit, P. Aungsuratana, and W. Khantiyanan, “Machine Learning Techniques for Short-Term Rain Forecasting System in the Northeastern Part of Thailand.” *Proceedings of World Academy of Science, Engineering and Technology, 31*, 2008, pp. 248 -253.

[9] L. Ingsrisawang, S, Ingsriswang, P. Luenam, P. Teesaranuwatana, and S. Somchit, *The Development of Rainmaking Forecast System over the Central and Eastern River Basin, Thailan(Final Report*). Kasetsart University and Bureau of Royal Rainmaking and Agricultural Aviation. Thailand, 2009.

[10] K.Y. Liang, and S.L. Zeger, “Longitudinal data analysis using generalized linear Models.” *Biometrika, 73*¸ 1986, pp. 13-22.

[11] P. McCullagh, and , J.A. Nelder. *Generalized Linear Models* (2nd ed.). London: Chapman &Hall, 1989.

[12] J. A. Nelder, and R.W.M. Wedderburn, “Generalized linear models.” *J.R. Statist. Soc. A 135*, 1972, pp. 370-384.

[13] W. Khantiyanan, *Analysis of GPCM forecasting model results.* Bureau of Royal Rainmaking and Agricultural Aviation. Thailand. 1996.

Table 1**:** Parameter estimates and its odds ratios from logistic models for prediction of P(W/W) and P(W/D) over the eastern Thailand**.**

Parameter	DF	Estimate	Standard Error	Wald Chisq	Pr >chisq	OR	95% CI of OR	
Prediction model for P(W/W)								
Intercept	1	0.2065	0.0278	55.2723	<.0001			
sin(2πt/K)	1	-0.4245	0.0306	192.4742	<.0001*	0.654	0.616	0.695
cos(2πt/K)	1	-0.7472	0.0408	334.7986	<.0001*	0.474	0.437	0.513
Prediction model for P(W/D)								
Intercept	1	-1.3392	0.0213	3961.75	<.0001			
sin(2πt/K)	1	-0.3184	0.0291	120.1129	<.0001*	0.727	0.687	0.77
cos(2πt/K)	1	-1.2593	0.0303	1729.0743	<.0001*	0.284	0.267	0.301
note: * p < 0.05								

Table 2. Final Multivariate GEE model for prediction the average of daily rain volume on wet days over the eastern Thailand**.**

Parameter	Estimate (β)	Standard error: SE (β)	Z	P-value	95% CI of β	
					Lower	Upper
Intercept	-5.2819	1.4790	-3.57	0.0004		
Royal rain operation						
yes	-0.1312	0.0828	-1.58	0.1133	-0.2935	0.0311
no (reference)	-	-				
The -15 °c isotherm height	0.0003	0.0001	4.43	<0.0001*	0.0001	0.0005
K-index	0.0456	0.0074	6.14	<0.0001*	0.0311	0.0601
Sweat- index	-0.0005	0.0012	-0.47	0.6379	-0.0029	0.0019
The average wind direction at the altitude of 20,000 -25,000 feet	-0.0007	0.0004	-1.58	0.1141	-0.0015	0.0001
note: * p < 0.05 , Deviance/df = 0.76, RMSE =10.29						