Jurnal
Teknologi

Full paper

Effect of Zero Measurements in Rainfall Data

Jamaludin Suhaila[a], Kong Ching-Yee[a*], Fadhilah Yusof[a], Foo Hui-Mean[a]

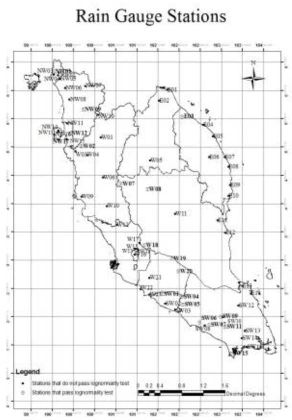
[a]*Department of Mathematical Sciences, Faculty of Science UniversitiTeknologi Malaysia,, 81310 UTM Johor Bahru, Johor, Malaysia*

*Corresponding author : suhailasj@utm.my

Article history

Received :21 January 2013
Received in revised form :
7 May 2013
Accepted :25 June 2013

Graphical abstract


Rain Gauge Stations
Legend
Stations that do not pass lognormality test
Stations that pass lognormality test
0 0.2 0.4 0.8 1.2 1.6
Decimal Degrees

Abstract

Flood is a commonly occurring hazard in Malaysia. The climate change in combination with the sea level rise will affected the frequency of flood events especially in a tropical country like Malaysia. Many researches focused on modeling rainfall data have been carried out in Malaysia. However, most of the rainfall studies did not include the zero values. The importance of these zero measurements should be examined in order to increase the quality of the research. The main purpose of this paper is to study the effect of zero measurement in rainfall analysis by applying a mixed bivariate lognormal distribution. The inter-station correlation coefficient was calculated in three cases of datasets. The first case considered only the positive values at both stations, and the second case included the positive values at either one of the stations, while the third case considered all values including zeroes at both rainfall stations. It was found that only the cases considering the positive measurements are useful and valid for the characterization of rainfall fields in our analysis.

Keywords: Mixed bivariate lognormal distribution; zero measurements; inter-station coefficient correlation

Abstrak

Banjir merupakan salah satu bencana alam yang sering melanda di Malaysia. Perubahan iklim dan kenaikan aras laut akan menjejas kekerapan banjir, terutama di negara tropika seperti Malaysia. Banyak kajian yang telah dijalankan di Malaysia memberi tumpuan kepada pemodelan hujan data. Walau bagaimanapun, kebanyakan kajian tersebut tidak merangkumi nilai sifar. Kepentingan nilai sifar perlu diselidiki bagi meningkatkan kualiti penyelidikan. Tujuan utama kertas kerja ini adalah untuk mengkaji kesan nilai sifar dalam kajian hujan dengan menggunakan taburan cantuman bivariat lognormal. Pekali korelasi antara stesen dihitung bagi tiga jenis data. Kes pertama mempertimbangkan hanya nilai-nilai positif di kedua-dua stesen, dan kes kedua adalah termasuk nilai-nilai positif di salah satu stesen, manakala kes ketiga mempertimbangkan semua nilai termasuk sifar di kedua-dua stesen hujan. Didapati bahawa hanya kes pertama yang merangkumi nilai positif adalah berguna dan sah dalam analisis ini.

Katakunci: Taburan cantuman bivariat lognormal; kesan nilai sifar; pekali korelasi antara stesen

■1.0 INTRODUCTION

Most of the rainfall studies conducted in Malaysia excluded the zero values from the analysis. Without the zero values, the distribution of rainfall will eventually become a continuous distribution. However, the rainfall characteristic is known to have a mixed property, which includes both discrete and continuous values. Zero values, which represent the non-rainy days, are considered as a discrete distribution, while nonzero rainfall values are considered as a continuous distribution. Since the mixed distribution includes the possibility of the no rain phenomenon and the skewness of real rain, the concept of mixed distribution is introduced in 1990's [1]. Based on the study, the mixed lognormal distribution was found to be an excellent fit to the average rain rate as well as having a tendency to provide an adequate model. Their findings have successfully increased awareness of zero values in rainfall studies.

Excluding the zero values without any appropriate test will put the validity of the studies at risk. Even when two nearby rain gauge stations are located within one river basin, the data collected from the two stations can be very different. When it is raining at station A, station B might not have rain at all. Two nearby rain gauge stations can have different fitted distributions [2]. In that sense, the total length of dry and wet periods and the characteristics of the wet and dry conditions become of interest in rainfall analyses such as trend analysis, fitting of distributions, and climate change studies. Even though the zero measurements are assumed to be important, they are not yet seriously considered in any of the studies in Malaysia. Zero values are assumed to be a barrier preventing easier characterization of rainfall in both time and space [3]. To investigate the importance of zero measurements in rainfall studies, two studies with a focus on the inter-station correlation coefficient with respect to the distance between the two rain gauges have been conducted [3-4]. They considered three possible cases of data structure in the problem. The first case considers only positive measurements at both rain gauge stations. The second case considers both zero and positive measurements, but either one of the two measurements has to be positive. The third case considers all the measurements including zero at both rain gauge stations. Bivariate mixed lognormal distributions have been applied to analyse these three possible cases [5]. The main finding of their works was that only the case which considers only positive values at both rainfall stations provides correlation estimates that are useful for the characterization of the rainfall fields.

Preliminary studies investigating the rainfall pattern and distributions in Malaysia have been successfully conducted by many researchers. For example, some studies concluded that the mixed distribution is found to be better than a single distribution in fitting rainfall data [6-8]. However, the analysis was done based on positive rainfall values where they only considered the rainfall amount on wet days without considering the zero values. The objective of this study is to determine the importance of zero values in rainfall analysis by applying a mixed bivariate distribution. The proposed distribution is lognormal. The present study is conducted in two ways: first, the analysis is done for stations that pass the lognormality test; second, the analysis is conducted for the stations that are based on the lognormal assumption.

■2.0 MATERIALS AND METHODS

2.1 Study Area and Data

The daily rainfall data were obtained from the Malaysian Meteorological Department and Drainage and Irrigation Department. The period of the study was from 1975 to 2007, a period of 33 years. The percentage of missing values during those periods was found to be less than 10% and they were estimated using the several weightings method [9]. The rainfall data were then checked through the homogeneity tests such as the standard normal homogeneity test, Buishand range test, the Pettit test, and the Von Neumann ratio test to ensure the quality of the data [10]. In total, 70 rain gauge stations were chosen after the homogeneity test. These stations were scattered over Peninsular Malaysia and their locations are shown in Figure 1. In this study, station NW17 is chosen as the target stations. The reason of choosing station NW17 is this station has many surrounding stations that nearly located.

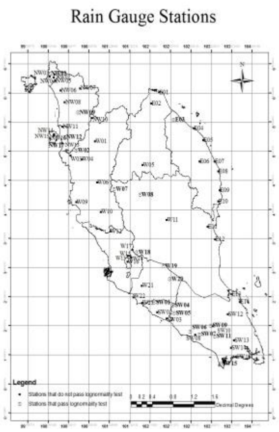


Figure 1 The locations of 70 stations in Peninsular Malaysia

2.2 Bivariate Model for Rainfall Data

The rainfall data used in this study can be categorized into four types of datasets: (0, 0), (*x**, 0), (0, *y**), and (*x*,*y*), where *x**, *y**, *x*, and *y* stand for positive values. The dataset is shown in Table 1, where $\sum_{r=0}^{3} n_r = N$, $x_i^* > 0 \ (i = 1,...,n_1)$, $y_j^* > 0 \ (j = 1,...,n_2)$, $x_k, y_k > 0 \ (k = 1,...,n_3)$, and the number $n_r (r = 0, 1, 2, 3)$ is a non-negative integer.

Table 1 Restructured rainfall data at two rain gauge stations

n_0	n_1	n_2	n_3
$0,\ldots,0$	$x_1^*,\ldots,x_{n_1}^*$	$0,\ldots,0$	$x_1,\ldots,x_{n_3}$
$0,\ldots,0$	$0,\ldots,0$	$y_1^*,\ldots,y_{n_2}^*$	$y_1,\ldots,y_{n_3}$

Let a non-negative random vector (X, Y) be the value of rainfall measurements at two rain gauge stations. The type of data in Table 1 suggested that a bivariate mixed distribution should be used to model the rainfall measured at two rain gauge stations [5]. The probability distribution of (X, Y) can be represented as:

$$
\begin{aligned}
&P(X = 0, Y = 0) = \delta_0 \\
&P(0 < X \le x, Y = 0) = \delta_1 F(x), \qquad x > 0 \\
&P(X = 0, 0 < Y \le y) = \delta_2 G(y), \qquad y > 0 \\
&P(0 < X \le x, 0 < Y \le y) = \delta_3 H(x, y), \qquad x, y > 0
\end{aligned} \tag{1}
$$

Where $0 \le \delta_r < 1 \, (r = 0,1,2,3)$ and $\delta_0 + \delta_1 + \delta_2 + \delta_3 = 1$, F and G are univariate positive continuous distribution functions, and H is a bivariate positive continuous joint distribution function. The conditional distribution at both of the rain gauge stations or at either one of the stations is as follows:

$$
\begin{aligned}
&P(X \le x \,|\, 0 < X \le x, Y = 0) = F(x) \qquad x > 0 \\
&P(Y \le y \,|\, X = 0, 0 < Y \le y) = G(x) \qquad y > 0 \\
&P(X \le x, Y \le y \,|\, X > 0, Y > 0) = H(x, y) \qquad x, y > 0
\end{aligned} \tag{2}
$$

In modelling rainfall data, any kind of positive skewed distribution can be used for *F* or *G* and any bivariate distribution such as an exponential, gamma, or lognormal joint distribution can be used for *H* [11].However, only 19 rainfall stations were found to pass the lognormality test, while the rest showed a different fitted distribution. However, based on the findings by

[2], the majority of the stations showed a mixed lognormal distribution or a univariate case. Therefore, the bivariate mixed lognormal distribution defined by [5] and [10] was adopted for the bivariate analysis in this study. The method of parameter estimation for bivariate mixed lognormal distribution can also be found in [5]. The estimated parameters were then substituted into the equations formulated by [3] and [4] to find the inter-station correlation coefficient between two stations.

2.3 Inter-station Correlation Coefficient Between Two Stations

In order to compute the inter-station correlation coefficient of rainfall data between two rainfall stations, this study follows the formulated equation simplified by [3] and [4]. The relationship between the restructured rainfall data of two rain gauge stations can be determined in the following three cases.

(i) Case A, where $A=\{X>0 \; and \; Y>0\}$ and the data (x,y) are used;

(ii) Case B, where $B=\{X>0 \; or \; Y>0\}$ and the data $(x^*,0)$, $(0,y^*)$, and (x,y) are used;

(iii) Case C, where $C=\{X\geq 0 \; and \; Y\geq 0\}$ and all of the data $(0,0)$, $(x^*,0)$, $(0,y^*)$, and (x,y) are used.

(iv) We denote ρ_i as the inter-station correlation coefficients under the three circumstances where $i=A,B,C$. The inter-station correlation coefficients ρ_A and ρ_B are conditional on A and B, while ρ_C is the unconditional inter-station correlation coefficient.

The well known theorem for conditional expectation by [12] is used to derive the relationship between the inter-station correlation coefficients. The theorem can be expressed as follows:

$$E\left[h(Y)\right]=\sum_{over\ x} E\left[h(Y)\,|\,X=x\right]P(X=x) \quad (3)$$

By applying the theorem, the relationship between the moment under the condition C can be denoted by $E\left(X^k\right)$, and the moment under the condition A or B can be denoted by $E\left(X^k\,|\,A\right)$ or $E\left(X^k\,|\,B\right)$ respectively. The following equations can be found in [4]:

$$\begin{aligned} E\left(X^k\right) &= E\left(X^k\,|\,B\right)P(B)+E\left(X^k\,|\,B^c\right)P\left(B^c\right) \\ &= \left(1-\delta_0\right)E\left(X^k\,|\,B\right) \end{aligned} \quad (4)$$

$$E\left(X^k\right)=\delta_1 E\left(X^k\,|\,X>0,Y=0\right)+\delta_3 E\left(X^k\,|\,A\right) \quad (5)$$

where $P(B)$ is the probability of B and $P(B)=1-P\left(B^c\right)$. Using the same theorem as that applied for Equations (4) and (5), the moments of Y are as follows:

$$\begin{aligned} E\left(Y^k\right) &= \delta_2 E\left(Y^k\,|\,X=0,Y>0\right)+\delta_3 E\left(Y^k\,|\,A\right) \\ &= \left(1-\delta_0\right)E\left(Y^k\,|\,B\right) \end{aligned} \quad (6)$$

$$E\left(X^kY^k\right)=\delta_3 E\left(X^kY^k\,|\,A\right)=\left(1-\delta_0\right)E\left(X^kY^k\,|\,B\right) \quad (7)$$

All equations used to find the values of ρ_A, ρ_B, and ρ_C can be found in [3] and [4].

■3.0 RESULTS AND DISCUSSION

A testing procedure for bivariate lognormality suggested by [11] was conducted for all 70 stations. The station NW17 was chosen as the target station along with another 18 neighbouring stations that passed the lognormality test based on the lowest value indicated by the AIC criterion. The distance between stations ranged from 21 to 533 km. The analysis in this study is conducted in two parts. First, the study will compute the inter-station correlation coefficients between the station NW17 and the other 69 neighboring stations based on the assumption of lognormality. Next, the inter-station correlation coefficients will be recomputed again with only those stations that passed the lognormality test.

Figure 2 displays the inter-station correlation coefficient of station NW17 with each of the other 69 stations based on the lognormal assumption, while the inter-station correlation coefficient between station NW17 and the 18 stations that pass the test of lognormality is shown in Figure 3. Although both of the figures shown different trend lines for the inter-station correlation coefficients of the three cases, some similarities do exist. From both figures, the inter-station correlation coefficients in case A are located near the zero regions; the inter-station correlation coefficients of case B are mostly negatively correlated, and the inter-station correlation coefficients of case C are usually positively correlated. Besides, the inter-station correlation coefficient of case C is the highest compared to case A and case B. The inter-station correlation coefficient of case B is the lowest among the three cases.

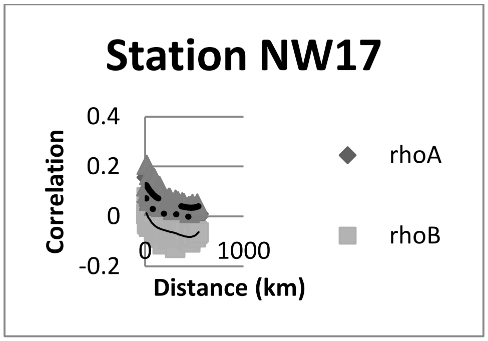


Figure 2 Spatial correlations estimated for the three cases of station NW17 based on the lognormal distribution assumption with the other 69 stations (rho is ρ)

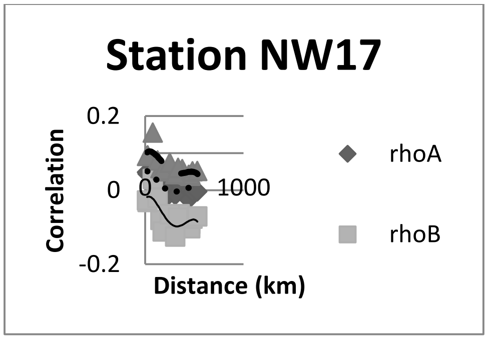


Figure 3 Spatial correlations estimated for the three cases of station NW17 with 18 stations that pass the test of lognormality (rho is ρ)

Cases A, B, and C in both figures provide consistent inter-station correlation coefficients with respect to the distance between gauges due to small variations. However, the inter-station correlation coefficients of cases A, B, and C in Figure 3 are slightly more stable than in Figure 2. The difference in the variability of the correlation coefficient between both figures is hard to determine from the graphical display. Hence, the variances were computed and are shown in Table 2. The lowest variability occurs for case A, followed by case C. The variability of case B is between those of case A and case C. The percentage variability for the inter-station correlation coefficient of case A is 0.09% for those stations which are based on the lognormal assumption, while for stations that passed the lognormality test it is around just 0.05%. The difference is nearly 0.04%. Similarly for case C, the percentage variability for stations that are based on lognormal assumption is greater than the variance for stations that passed the lognormality test. Overall, the result is more accurate for those stations that have been proven to follow the lognormal distribution than for those for which the lognormal assumption is used. Based on the analysis, the study found that the value of the inter-station correlation coefficient for case A is the most consistent, compared to the other two cases. It is shown that case A, which only includes the positive values from both stations, is useful and valid for the characterization of rainfall fields in our analysis.

Table 2 The percentage variance of the inter-station correlation coefficients of cases A, B and C

Inter-station correlation coefficient	ρ_A	ρ_B	ρ_C
Stations with lognormal assumption	0.09%	0.14%	0.13%
Stations that passed the lognormality test	0.05%	0.14%	0.08%

■4.0 CONCLUSION

The importance of zero values in rainfall research was analysed in this study. A bivariate mixed distribution is used to compute the inter-station correlation coefficient for four types of datasets. Three cases were considered: Case A, which considered positive measurements from both stations, Case B, which considered the positive measurements from either one or both stations, and Case C, which considered all the rainfall measurements including zeroes at both of the stations.

The research was carried out using station NW17 as the target station. The analysis has been done using two different approaches: firstly, the inter-station correlation coefficient was computed for all stations based on the lognormality assumption, and secondly the inter-station correlation coefficient was computed only for stations that passed the test of lognormality. More variability was observed when considering the zero measurements (cases B and C) compared to case A. Case A showed the lowest variability, especially for those stations that have been proven to follow a lognormal distribution.

Acknowledgement

This research was financed by MyPhD and FRGS vote 4F024 from the Ministry of Higher Education of Malaysia. The authors are grateful to the Malaysia Meteorological Department and Drainage and Irrigation Department for providing the daily rainfall data as the research material.

References

[1] B. Kedem, L. S. Chiu, and G. R. North. 1990. Estimation of Mean Rain Rate: Application to Satellite Observations. *Journal of Geophysical Research*. 95**:** 1965–1972.

[2] J. Suhaila, C.-Y. Kong, Y. Fadhilah, and H.-M. Foo. 2011. Introducing the Mixed Distribution in Fitting Rainfall Data. *Open Journal of Modern Hydrology*. 1**:** 11–22.

[3] C. Yoo, and E. Ha. 2007. Effect of Zero Measurements on the Spatial Correlation Structure of Rainfall. *Stochastic Environmental Research and Risk Assessment*. 21**:** 287–297.

[4] E. Ha, and C. Yoo. 2007. Use of Mixed Bivariate Distributions for Deriving Inter-station Correlation Coefficients of Rain Rate. *Hydrological Processes*. 21**:** 3078–3086,

[5] K. Shimizu. 1993. A Bivariate Mixed Lognormal Distribution with an Analysis of Rainfall Data. *Journal of Applied Meteorology*. 32(2)**:** 161–171.

[6] J. Suhaila, and A. A. Jemain. 2007. Fitting Daily Rainfall Amount in Peninsular Malaysia Using Several Types of Exponential Distributions. *Journal of Applied Sciences Research*. 3(10): 1027–1036.

[7] J. Suhaila, and A. A. Jemain. 2009. Investigating the Impacts of Adjoining Wet Days on the Distribution of Daily Rainfall Amounts in Peninsular Malaysia. *Journal of Hydrology*. 368:17–25.

[8] S. M. Deni, J. Suhaila, W. Z. W. Zin, and A. A. Jemain. 2010. Spatial Trends of Dry Spells Over Peninsular Malaysia During Monsoon Seasons. *Theoretical and Applied Climatology*. 99: 357–371.

[9] J. Suhaila, M. D. Sayang, and A. A. Jemain. 2008. Revised Spatial Weighting Methods for Estimation of Missing Rainfall Data. *Asia-Pacific Journal of Atmospheric Sciences.* 44**:** 93–104.

[10] J. B. Wijngaard, A. M. G. Klein Tank, and G. P. Können. 2003. Homogeneity of 20th century European daily temperature and precipitation series. *International Journal of Climatology*. 23**:** 679–692.

[11] K. Shimizu, and M. Sagae. 1990. Modeling Bivariate Data Containing Zeros, with an Analysis of Daily Rainfall Data. *Japanese Journal of Applied Statistics*. 19: 19–31.

[12] E. Parzen. 1962. *Stochastic Processes.* San Francisco, Holden-Day.