Fitting monthly Peninsula Malaysian rainfall using Tweedie distribution

Fitting monthly Peninsula Malaysian rainfall using Tweedie distribution

R M Yunus[1], M M Hasan and Y Z Zubairi[2]

[1]Institute of Mathematical Sciences, University of Malaya, Kuala Lumpur, Malaysia
[2]Centre for Foundation Studies in Sciences, University of Malaya, Kuala Lumpur, Malaysia

rossita@um.edu.my

Abstract. In this study, the Tweedie distribution was used to fit the monthly rainfall data from 24 monitoring stations of Peninsula Malaysia for the period from January, 2008 to April, 2015. The aim of the study is to determine whether the distributions within the Tweedie family fit well the monthly Malaysian rainfall data. Within the Tweedie family, the gamma distribution is generally used for fitting the rainfall totals, however the Poisson-gamma distribution is more useful to describe two important features of rainfall pattern, which are the occurrences (dry months) and the amount (wet months). First, the appropriate distribution of the monthly rainfall was identified within the Tweedie family for each station. Then, the Tweedie Generalised Linear Model (GLM) with no explanatory variable was used to model the monthly rainfall data. Graphical representation was used to assess model appropriateness. The QQ plots of quantile residuals show that the Tweedie models fit the monthly rainfall data better for majority of the stations in the west coast and mid land than those in the east coast of Peninsula. This significant finding suggests that the best fitted distribution depends on the geographical location of the monitoring station. In this paper, a simple model is developed for generating synthetic rainfall data for use in various areas, including agriculture and irrigation. We have showed that the data that were simulated using the Tweedie distribution have fairly similar frequency histogram to that of the actual data. Both the mean number of rainfall events and mean amount of rain for a month were estimated simultaneously for the case that the Poisson gamma distribution fits the data reasonably well. Thus, this work complements previous studies that fit the rainfall amount and the occurrence of rainfall events separately, each to a different distribution.

1. Introduction

Fitting the rainfall data using the most suitable probability distribution helps in improving the quality of predicting rainfall characteristics, for example the mean amount of rainfall per occurrence, the number of rainfall events, and the probability of no rain. In the literature, many studies use gamma distribution for modelling the amount of rainfall [1]-[4], and completely ignore the occurrence of rainfall events, or replace zero monthly rainfall values by a very small value such as 0.01 before fitting the data. The amounts of rainfall data fit well various distributions, namely the gamma, normal, Weibull, and log normal [5], depending on the geographic location of the studied stations. Since rainfall variables are naturally consist of both the discrete and continuous components, some researchers used two separate models to model the occurrence and amount of rainfall [6]-[8]. In recent studies, the Tweedie distributions fit well the monthly rainfall data from stations all over Australia [9]-[10], and describe both the amount of rainfall and the occurrence of rainfall events, simultaneously, thus complement the works of previous studies.

In this paper, we aim at fitting the monthly rainfall data in Malaysia using the distribution within the Tweedie family. The studies for modelling monthly rainfall data in Malaysia are still exhaustive. Many studies in the literature are focusing on fitting the daily rainfall [11]-[13], for example Yunus et al. [14] used the Tweedie model to model the daily rainfall of four stations in Peninsula Malaysia. Some studies on monthly rainfall of Malaysia find precipitation index, to evaluate the dry and wet spells, and to model the frequencies transition of wet categories [15]-[16]. These studies however, did not aim to find the distribution that best fit the monthly rainfall data in Malaysia.

In section 2, the rainfall data and Tweedie model are described. In section 3, the adequacy of the fitted model is validated through observing the QQ plot of quantile residuals. The Tweedie generated rainfalls frequency histogram and that of the observed (actual) rainfall are compared through simulation, in the same section. Conclusion is given in the final section.

2. Material and Method

In this section, data and method are explained.

2.1. Data

Rainfall data are obtained from the Malaysia Meteorological Department. Figure 1 shows the location of the studied stations. The stations that are located in the west region of Peninsula Malaysia receive southwest monsoon from late May to September. A spine of mountain ranges running down the centre of Peninsula Malaysia, from north to south, divides the east and west side into two monsoon regimes. The high lands prohibit the west side from receiving the northeast monsoon that normally bring heavy rainfall, particularly to the east coast of Peninsula Malaysia. Table 1 shows descriptive statistics of monthly rainfall data from January, 2008 to April, 2015 for stations with some month without rain.

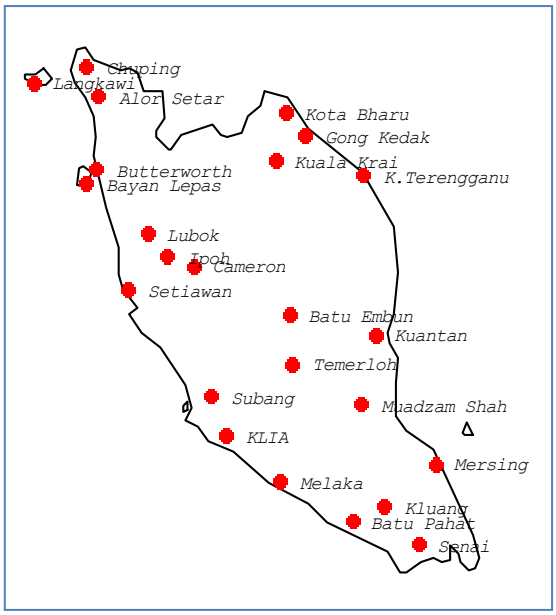

Figure 1. Location of studied stations.

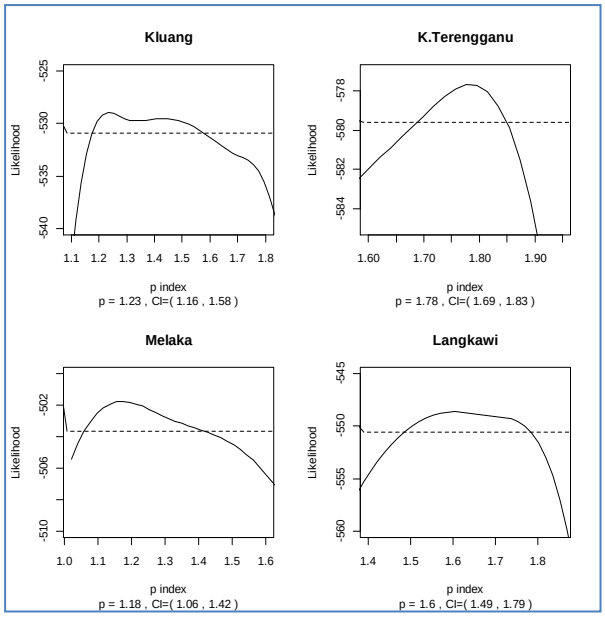

Figure 2. The profile likelihood plot for the studied stations used to find the MLE of p.

2.2. Tweedie disribution

The Tweedie distribution is defined by three parameters, mean μ, dispersion parameter ϕ and index parameter p. Tweedie family of distributions are normal, when $p = 0$; Poisson, when ($p = 1$ and $\phi = 1$) ; gamma when $p = 2$; inverse Gaussian, when $p = 3$; Poisson gamma when $1< p < 2$; and a positive right skew distribution when $p \geq 2$. Poisson gamma variable is positive continuous with exact

zeroes, so this distribution is suitable for modelling the rainfall data with exact zeros. However, its probability distribution function is not a closed form.

Table 1. Descriptive statistics of monthly rainfall from 2008 to 2015 for selected stations.

	Kluang	K.Terengganu	Melaka	Langkawi
Minimum (mm)	0.00	0.00	0.0	0.0
1st Quartile (mm)	89.62	90.05	89.9	55.8
Median (mm)	184.00	148.80	143.4	198.1
Mean (mm)	183.81	254.61	153.4	198.5
3rd Quartile (mm)	243.40	314.85	211.8	280.1
Maximum (mm)	508.40	1580.40	364.0	871.0
Number of months with no rain	1	1	1	2

In the initial studies, GLM is proposed to fit a model when the observations of the response variables were not from a normal distribution [17]. GLMs allow the response variable to have a probability distribution which is coming from the class of the EDM family of distributions. The Tweedie distribution is a special case of EDMs family [18].

The modelling of monthly rainfall using this model is adopted from Hasan and Dunn [9]-[10]. We assume any rainfall event i produces an amount of rainfall R_i for $i = 1,2, \dots, N$, where R_i comes from a gamma distribution with mean $\alpha\gamma$ and variance $\alpha\gamma^2$. We also assume the number of rainfall events in any one month is N, where N has a Poisson distribution with mean λ. This means $N = 0$ denote the months with no rainfall. The total monthly rainfall $Y = R_1 + R_2 + \cdots + R_N$. Then, the Tweedie random variable Y has mean μ and variance $\phi\mu^p$ [19]-[22] using some functional relationship between the response and linear predictor of GLM.

Based on GLM, the rainfall data is fitted using

$$\log \mu_i = \beta, \quad (1)$$

where μ_i is the expected value of the Tweedie variable Y_i and β, is a regression coefficient, $i = 1, 2, \dots, 88$ (for the period from Jan, 2008 to April, 2015).

In order to fit the Tweedie distribution on rainfall data, we need to estimate parameters β, ϕ and p. The estimate of β is obtained using some iterative procedure, through numerical computation, in particular function 'glm' with Tweedie family in R. The profile likelihood plot is used to estimate p, and it is the value for which the log likelihood is maximized. The algorithms in the 'tweedie.profile' function in R package 'tweedie' [22] estimate p and also ϕ.

3. Results and Discussion

The initial step in fitting the Tweedie model on monthly rainfall data is estimating the p index for all stations. From the profile plots given in figure 2, the MLEs of p are 1.23, 1.78, 1.18 and 1.60 for Kluang, Kuala Terengganu, Melaka and Langkawi stations, respectively. Thus, the distribution of monthly rainfall for these stations is Poisson gamma. For stations with no month without rain, the distribution of the monthly rainfall for these stations is gamma ($p = 2$).

The MLEs of p are then used to obtain the QQ plot of the quantile residuals, as a diagnostic tool to check the adequacy of the fitted model. Figure 3 depicts that all the points lie on or close to the straight line for majority of the stations, suggesting that the Tweedie distributions are sufficiently adequate for monthly rainfall for these stations. No large deviations were observed except at the upper tail of the QQ plot for most of the stations in the east coast of Peninsula, namely stations Gong Kedak, Kota Bharu, Kuala Krai, Kuala Terengganu, Kuantan and Muadzam Shah. The plots suggest that the Tweedie distribution fit reasonably well for most stations in the west coast and midland (except stations Ipoh and Cameron) than those in the east coast of Peninsula. The distribution of the monthly rainfall of a station is very much dependent on the geographical location of the station.

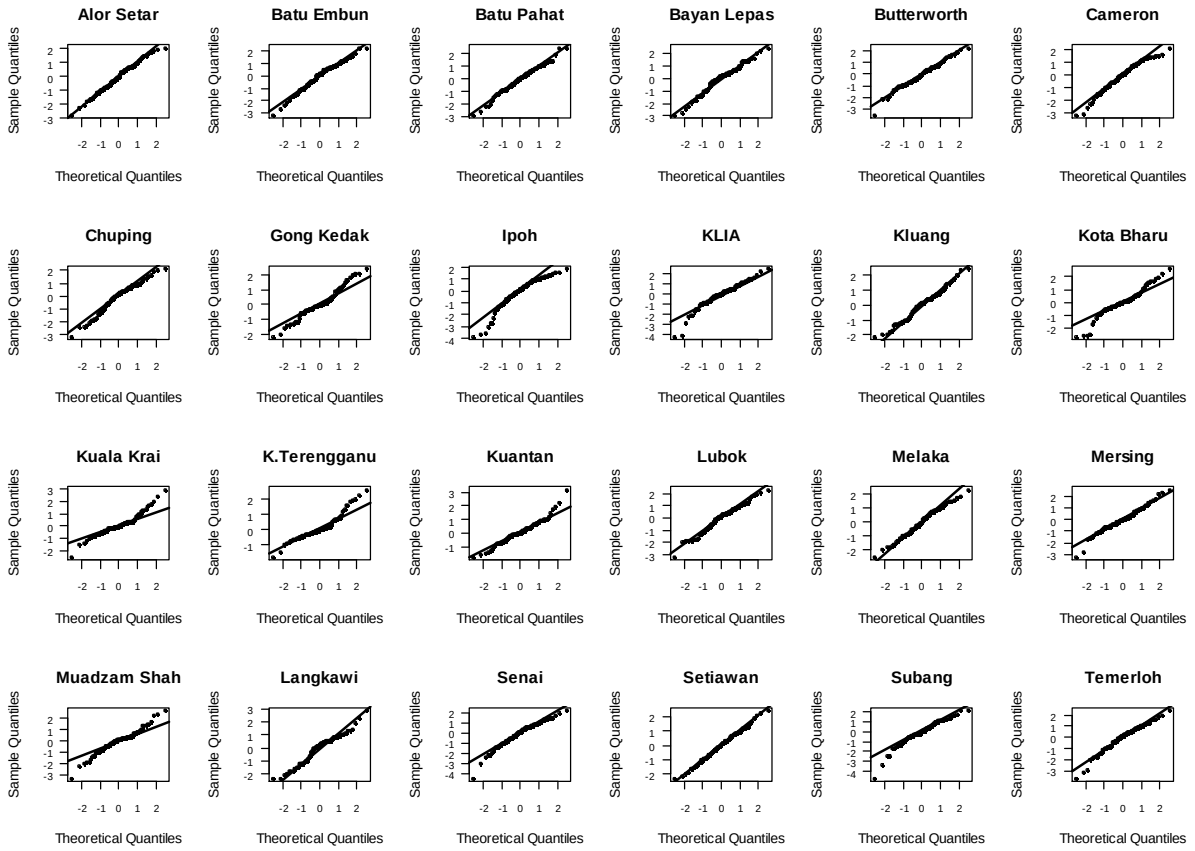

Figure 3. QQ plot of quantile residuals from the model.

In this paper, we have developed a simple model for simulating rainfall for use of various purposes in various studies. In this section, simulated rainfall samples of size 100 were generated from Tweedie EDM with μ, p and ϕ are the respective MLEs for each studied stations. Frequency histograms of simulated dataset as well as the actual dataset were plotted for each station, to determine whether the simulated data have a similar distribution as the observed (actual) rainfall data. From figure 4, we find that the simulated data using the Tweedie model has fairly similar frequency histogram as the actual data, suggesting the Tweedie models fit well the monthly rainfall data for most of the stations. Accordingly, the Tweedie model can be used to generate synthetic rainfall data, and to fill missing gaps to complete the inadequate observed rainfall records.

The fitted Poisson gamma model can be used to study various features of monthly rainfall, namely the mean number of rainfall events, λ, the shape of the rainfall gamma distribution, γ, and the amount of rain per rainfall event, $\alpha\gamma$, and probability of no rain, π, through the following relationship [23]

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}; \qquad \gamma = \phi(p-1)\mu^{p-1}; \qquad \alpha = \frac{p-2}{1-p}.$$

For example, for Langkawi St., the MLE of ϕ is 5.26, the mean number of events for a month is estimated as $\hat{\lambda} = 3.74$, the shape of the gamma distribution $\hat{\gamma} = 86.71$ and the mean amount of rain per event, $\hat{\alpha}\hat{\gamma} = 53.14$ mm.

It is of interest to know how well the model predicts the probability of no rain $\pi = \exp(-\lambda)$ [24]. The probability of no rain $\hat{\pi}$= 0.024 for station Langkawi. There are eighty eight months from January, 2008 to April, 2015. As a result, the estimated number of months without rain is $88 \times 0.024 \approx 2$ in the period study, and this is in agreement to the observed number of months with no rain that is given in table 1.

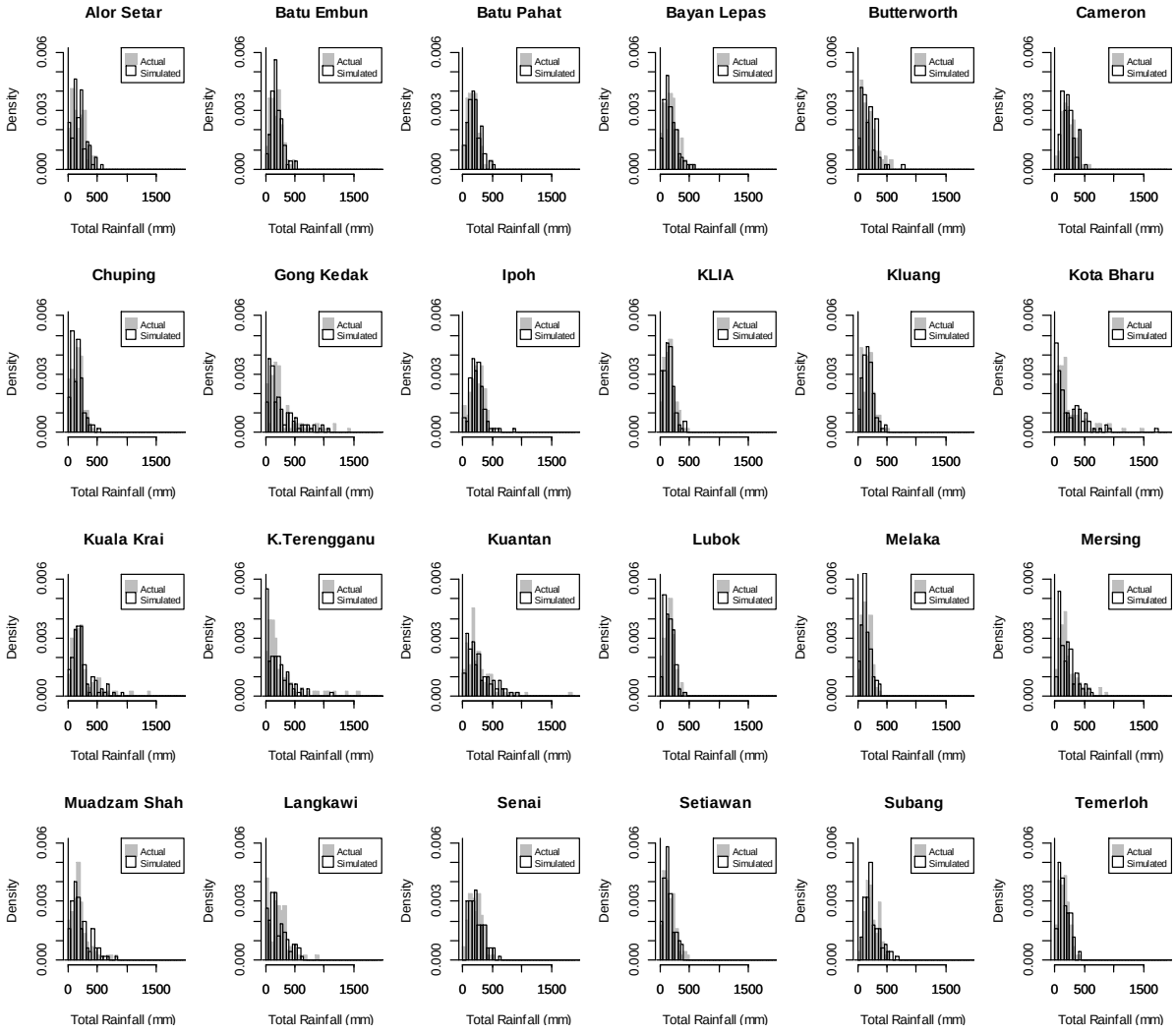

Figure 4. Histogram of the observed and simulated rainfall distributions.

4. Concluding Remarks

The time series Malaysian rainfall data have zero amounts of rainfall for some of the months in the studied period. Monthly rainfall data have two important features – the amount (wet) of rainfall and the occurrence (dry) of rainfall events. The search for the best fitting distribution for rainfall amount and occurrence of rainfall events has been the main interest in several studies across the world. Various forms of distributions have been examined to identify the best fitted distribution for the rainfall amount and the occurrence of rainfall events, and most of the previous works studied the two features separately.

The gamma and Poisson gamma are two distributions within the Tweedie family of distributions. Poisson gamma variable is a positive continuous with exact zeroes, thus the Poisson gamma model describes both the amount of rainfall and the occurrence of rainfall events, simultaneously. In this study, the Tweedie distributions were used on the monthly rainfall data, and thus complement the works of previous studies.

The Tweedie distribution was identified as fairly well at fitting monthly rainfall data with exact zeroes (i.e. no rain in a month) for majority of the stations in Peninsula. Based on the QQ plots of quantile residuals, it was identified that the monthly rainfall for stations in the west coast and midland

of Peninsula Malaysia, seems to fit the Tweedie distribution better than the east coast of Peninsula Malaysia. Thus, the distribution that best fit the monthly rainfall data of a station depends on the geographical location of the station.

The simple Tweedie model that is developed in this paper is useful for simulating rainfall data. The histograms of the simulated data were fairly similar to the actual data for most of the stations in Peninsula. This finding is important to generate synthetic rainfall data for stations with inadequate records.

The models are potentially important to simulate various characteristics of rainfall, such as the mean monthly rainfall amount, the probability of no rain and the number of rainfall occurrence per month, which may have potential uses in various areas including agriculture, water resource and irrigation. Despite explain characteristics of rainfall, the model fits the data well and produces reasonable simulated data, and it is therefore recommended for future use.

Reference

[1] Mooley DA 1973 *Weather Review* **101** 160–76
[2] Wilks DS 1999 *Agricultural and Forest Meteorology* **93** 153–69
[3] Chapman T 1998 *Environmental Modelling and Software* **13** 317–24
[4] Stern RD and Coe R 1984 *Journal of the Royal Statistical Society, Series A.* **147** 1–34
[5] Abtew W, Melesse AM and Dessalegne T 2009 *Hydrological Processes* **23** 3075–82
[6] Chandler RE and Wheater HS 2002 *Water Resources Research* **38** 1192–202
[7] Hamlin MJ and Rees DH 1987 *Hydrological Sciences* **32** 15–29
[8] Buishand TA, Shabalova MV and Brandsma T 2004 *Journal of Climate* **17** 1816–27
[9] Hasan MM and Dunn PK 2010 *Agricultural and Forest methodology* **150** 1319–30
[10] Hasan MM and Dunn PK 2011 *International Journal of Climatology* **31** 1389–97
[11] Suhaila J and Jemain AA 2007 *Journal of Applied Sciences* **7** 1800-86
[12] Suhaila J and Jemain AA 2007 *Journal of Applied Sciences Research* **3** 1027-36
[13] Deni SM, Jemain AA and Ibrahim K 2010 *International Journal Of Climatology* **30** 1194–205
[14] Wan Zin WZ, Jemain AA and Ibrahim K 2013 *Theor. Appl. Climatol.* **111** 559–68
[15] Yunus RM, Hasan MM, Razak NA and Zubairi YZ 2017 *International Journal of Climatology* **37** 1391-1399
[15] Sanusi W and Ibrahim K 2012 *Sains Malaysiana* **41** 1345–53
[16] McCullagh P and Nelder JA 1989 *Generalized Linear Models, 2nd edition* (London - Chapman and Hall)
[17] Jørgensen B 1987 *Journal of the Royal Statistical Society, Series B* **49** 127–62
[18] Jørgensen B 1997 *The Theory of Dispersion Models* (London - Chapman and Hall)
[19] Tweedie MCK 1984 *Proc. of the Indian Statistical Institute Golden Jubilee International Conference* pp.579-504 (Calcutta/Indian Statistical Institute)
[20] Smyth GK 1996 *Proc. of the Second Australia–Japan Workshop on Stochastic Models in Engineering, Technology and Management* pp. 572–80 (Brisbane/University of Queensland Technology Management Centre)
[21] Smyth GK with contributions from Hu Y and Dunn PK 2009 Statmod: Statistical Modeling R Package Version 1.4.1
[22] Dunn PK 2009 Tweedie: Tweedie Exponential Family Models R Package, Vienna, Austria, R Package Version 2.0.2
[23] Dunn PK 2004 *International Journal of Climatology* **24** 1231–1239
[24] Dunn PK and Smyth GK 2005 *Statistics and Computing* **15** 267–280

Acknowledgments

This work received financial support from University of Malaya Research Grant RG369-15AFR.