MATEMATIKA, 2013, Volume 29, Number 1b, 53-65

On the Applicability of Bartlett Lewis Model: With Reference to Missing Data

[1]Ibrahim Suliman Hanaish, [2]Kamarulzaman Ibrahim and [3]Abdul Aziz Jemain
[1]Department of Statistics, Faculty of Science, Misurata University, Misurata, Libya
[2,3]School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia,
[2]Solar Energy Research Institute, Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor, Malaysia
e-mail: [1]henaish@yahoo.com

Abstract The availability of a complete hourly rainfall dataset is required for hydrological applications, statistical modeling and forecasting of precipitation. The issues of missing data are of serious concern in rainfall modelling due to the problem in computing the autocorrelation when gaps of missing data are encountered in the pooled data. The present paper discusses the applicability of three methods of handling missing hourly data when the Bartlett Lewis rectangular pulses model is utilized: zero substitution, single imputation and multiple imputations. The three methods are applied to the hourly rainfall data from the Bukit Bendera rain gauge station, which consists of a complete nine year rainfall series. The methods are tested with different percentages of randomly generated missing rainfall values. The performance of the methods is studied in terms of the mean absolute deviation errors that are found during different monsoon periods. The findings indicate that the best method to address missing data when applying the Bartlett Lewis rectangular pulses model is the single imputation method.

Keywords Single Imputation; Multiple Imputations; Bartlett Lewis Rectangular Pulses Model

2010 Mathematics Subject Classification 91B70

1 Introduction

Missing values are a common problem faced during the analysis of time series data. A complete hourly rainfall dataset is required for many purposes, such as hydrological applications, statistical modelling and the forecasting of precipitation. Certain issues arise in modelling when missing data is encountered in the pooled data, particularly in the case of computing autocorrelations. There are several reasons for missing data. For example, hourly rainfall may be missing for a certain period of time because the rain gauge is not functioning. In addition, rainfall data may be treated as missing data where reporting errors, recording errors intermittent reporting are involved.

The hourly rainfall data have the characteristics of being highly spatial discontinuous and uncorrelated over time. The discrete nature of rainfall in time and space poses a unique problem for meteorologists and climatologists in comparison with more continuous variables, such as temperature and pressure. Therefore, most traditional approaches of infill are not suitable for hourly rainfall data. However, the estimation of hourly rainfall missing data remains a challenge that is rarely reported or addressed in extant literature.

The Malaysian National Network System (MNNS) uses three different methods for the collection of rainfall data. Most rainfall stations in Peninsular Malaysia apply these three methods concurrently, which inadvertently results in missing rainfall data. The missing data can be observed based in at least one of these recording methods at any given time. A significant difference in the data measurements collected from individual stations due to the use of different recording methods at the same time at individual

stations. Through a data exploration exercise, the discrepancy between one method of measurement and another are found to range between 0% - 100%, as mentioned by [1], which indicates the relative instability of the data.

The handling of missing data is an issue that is widely examined in extant literature regarding different daily rainfall data from various climates, but rarely for hourly rainfall data. Several methods exist for imputing missing rainfall data. The normal ratio method and the inverse distance weighting method are two commonly used methods, for example, [2] , [3] and [4]. Furthermore, [1] apply a model based on the combination of the artificial neural network (ANN) and the nearest neighbour imputation (NNeigh) techniques.

Tang [5] demonstrates that the spatial correlations of hourly rainfall amount between two rainfall stations in Peninsular Malaysia vary from month to month. Thus, the NNeigh method is commonly applied in Peninsular, which involves replacing missing data for a particular hour with the observed value of the nearest station. Additionally, the method is used by [6] for predicting the values of missing hourly rainfall data. In addition, [7] use the Kriging method for the spatial prediction of hourly rainfall data. Although the empirical orthogonal function method is the most effective tool in dealing with spatially inhomogeneous climate fields, the method may yield unreasonable results when the field is highly non-stationary, as in the case of hourly precipitation data [6]. The requirement of stationarity makes it a poor choice for the handling of missing hourly precipitation data [8].

Methods of handling missing data are effective if one is aware of the types of missing data. The three types of missing data are missing completely at random (MCAR); missing at random (MAR); and not missing at random (NMAR). Malaysian hourly rainfall missing data are considered to fall into the category of MCAR, which means that a particular pattern of missing data is not dependent on the values that are missing and is not dependent on the observed data [9].

In Peninsular Malaysia, a limited number of rain gauge stations exist that complete hourly rainfall observations over long periods. The present study presents some imputation methods for estimating the missing hourly rainfall data. While a number of extant works consider the handling of daily missing rainfall data in Peninsular Malaysia, few consider the manner in which to handle hourly missing rainfall data. Three methods of handling missing data are considered in the present study: zero substitution, single imputation and multiple imputations. The objective of this study is to examine the effects of the different methods of filling in missing values on the estimated rainfall statistics produced by the Bartlett Lewis model and to determine which of the three methods for handling missing rainfall data is best for rainfall modelling.

2 Stochastic Rainfall Modelling

Stochastic rectangular pulses cluster rainfall models are frequently applied in the field of hydrology. For example, such models can be used to generate rainfall data across a range of timescales for the purpose of reservoir design, flood studies and design of sewerage systems. The Bartlett-Lewis model, which is a popular cluster Poisson model, is based on stochastic cluster rainfall point processes models is found to accurately reproduce conventional rainfall statistics over a range of aggregation levels, normally between 1h and 1 day [10]. Many studies have considered such models for describing the rainfall patterns for a single site. Rodriquez-Iturbe et al. [11], [12], [13], [14-16] and [17] apply the Bartlett-Lewis model in their respective research. The Bartlett Lewis model is successful in describing the rainfall processes for a wide range of temporal scales in a variety of countries, such as the UK, the USA and Australia.

2.1 Modified Bartlett Lewis Model (MBL)

The modified Bartlett Lewis (MBL) model assumes that storms arrive following a Poisson process with rate λ. Within each storm, rectangular cells arrive following another Poisson process with rate β. The

duration of the storm is distributed according to an exponential distribution with parameter γ. The depth of a cell is assumed to follow an exponential distribution with parameter $1/\mu_x$ and the cell duration η is distributed following a gamma distribution with shape parameter α and scale parameter ν. Both β and γ are scaled with respect to the cell duration to obtain the dimensionless parameters $\kappa = \beta/\eta$ and $\phi = \gamma/\eta$. The MBL is thus defined by six parameters $(\lambda, \mu_x, \alpha, \nu, \kappa, \phi)$ as depicted in the schematic provided in Figure 1. The equations of the MBL model are presented within numerous extant studies [11]. The equations relate the statistical properties of the rainfall process in discrete time throughout the entire time domain to the model parameters and serve as the basis for model fitting.

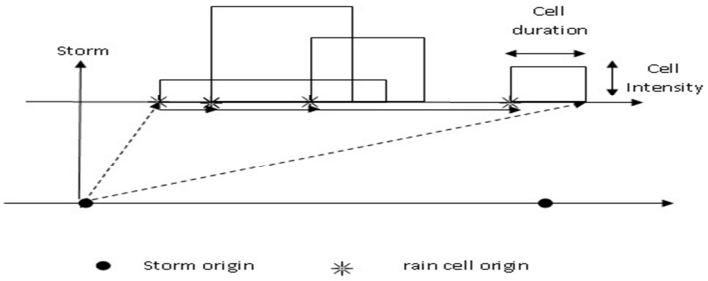

Figure 1 Explanatory sketch of the Bartlett-Lewis rectangular pulses model

2.2 Model Fitting

In the absence of any more sophisticated techniques for parameter estimation, the Nelder-Mead optimization algorithm is applied to minimize the objective function, which is as follows:

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{k} w_i \left[\left(1 - \frac{T_i(y)}{\tau_i(\theta)} \right)^2 + \left(1 - \frac{\tau_i(\theta)}{T_i(y)} \right)^2 \right] \tag{1}$$

where $\mathbf{T}(y) = (T_1(y), T_2(y), ..., T_k(y))$ is a vector of summary statistics computed based on the data y, $\tau(\boldsymbol{\theta}) = (\tau_1(\theta), \tau_2(\theta), ..., \tau_k(\theta))$ denotes a vector of the fitted value of $\mathbf{T}$ under the model and w_i is a collection of positive weights to avoid bias due to the differing orders of magnitudes of observed values of statistics involved which are assumed as $1/Var(T_i(y))$, where $Var(T_i(y))$ represents the i^{th} diagonal elements of the covariance matrix of the summary statistics. The sample moments used to determine model parameters are 1-hour mean, 24-hour mean, 1-hour variance, 24-hours variance, 1-hour lag-1 autocorrelation and 1-hour probability of wet.

2.2.1 The Nelder-Mead Algorithm

The Nelder-Mead algorithm is one of the most well-known algorithms [18] is widely used for parameter estimation and for problems with discontinuous functions which occur frequently in statistics and mathematical experiments. The popularity of the algorithm stems from its simplicity and ease of application. The algorithm is designed to resolve the classical unconstrained optimization problem of minimizing a given nonlinear function. Spendley et al. [19] introduce the basic formulation of the

algorithm, which is based on tracking ideal operating conditions by evaluating the output of a system at a set of points; forming a simplex in the parameter space; and continuously forming new simplices by reflecting a point in the hyperspace of the other points. The concept is later acknowledged by [18] as an optimized mathematical formulas. The formulation and the configuration of the Nelder-Mead algorithm is presented in greater detail in extant literature, such as [18] and [20].

3 Rainfall Data

The hourly data used in the present analysis are obtained from the Bukit Bendera rain gauge station in Peninsular Malaysia, which is situated at 5.42° N and 100.27° E as shown in Figure 2. Bukit Bendera is selected as a representative climatic data station due to the availability of a complete hourly rainfall data over a larger period of time when compared against other rain gauge stations. Rainfall varies seasonally at the Bukit Bendera station due to the occurrence of the monsoon winds. The seasonal variation is primarily influenced by the southwest monsoon season, which occurs between May and August, and the northeast monsoon season, which occurs between November and February. During the northeast monsoon season, many areas on the east coast of Peninsular Malaysia are expected to receive heavy rainfall. On the other hand, areas on the west coast that are sheltered by the mountain ranges are more or less free from the influence of the north east monsoon season. In addition, the transition period between the monsoons, i.e. the inter-monsoon period, occurs between the months of March to April and September to October. The hourly data are collected from the database of Malaysian Meteorological Service (MMS) and contain observations of the period between 1977 and 1986. One important feature of rainfall data in Malaysia must be considered when imputing missing values. The nature of rainfall of Malaysia results in a high percentage of dry hours and, consequently, a large percentage of the hourly rainfall values are 0. This feature of the data needs to be taken into account when imputing missing values. Figure 3, below, shows a histogram of the Bukit Bendera hourly rainfall measurements of the same four months of all nine years examined, during both of the aforementioned monsoon seasons and both inter-monsoon periods. The plots indicate a large proportion of zeros, as well as high proportion of the lowest depths of the number of rainfall hours for the months considered.

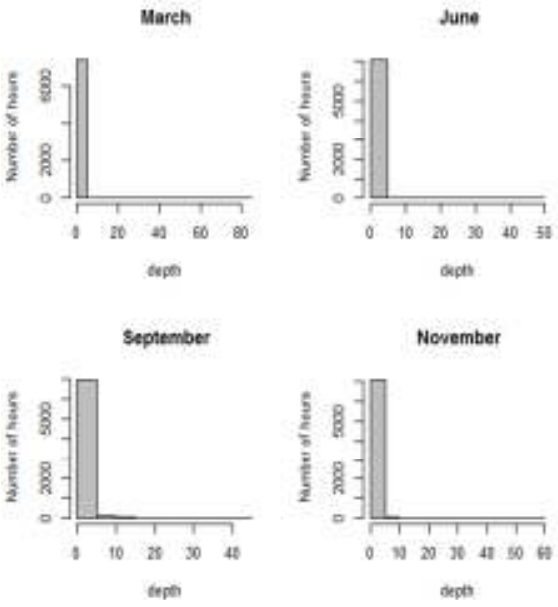

Figure.3 Hourly rainfall data for different months at Bukit Bendera station

Figure.2 geographical location of Bukit Bendera station

4 Missing Data Methods

4.1 Zero Substitution Method (ZS)

The zero substitution (ZS) approach is an ad hoc method for handling missing data in which the missing data are replaced with the zero values. The zero value is assumed to be the best estimate for any observation with a missing value. The limitation of the method is that it fails to take into account the monsoon periods and dry months, causing overestimation and underestimation of the rainfall amounts for the dry months and monsoon periods, respectively.

4.2 Hot-Deck Method

Several methods involving the imputation of missing data involve single imputation. A common single imputation method often utilized by researchers is the Hot-Deck method, which involves replacing each missing value with a randomly drawn value from the set of data values found within the same month [21]. Although this method is simple to employ, the drawback is that the variance estimates are almost always reduced [22] and [23] and do not reflect the uncertainty due to imputation. In the present approach, the imputation method utilized simulates the missing data by statistical distribution, which is assumed to be exponential based upon [24] and [25]. The method is referred to as a single imputation method (SI).

4.2.1 Single Imputation method (SI)

The first step of the single imputation method (SI) involves generating the occurrence of missing data for wet and dry hours. To determine if a particular hour is wet or dry, a uniform distribution on the interval $[0,1]$ is generated for a random variable, say r, and compared to the probability of wet (*Pwet)* of that particular month. If r is less than *Pwet*, then the hour is a dry hour. Otherwise, the hour is classified as a wet hour. Finally, the imputed values for the missing hours are drawn from a distribution which is assumed to be an exponential distribution with a rate λ which is equal to $1/Mn$ where Mn is the mean amount of wet hours for that particular month.

4.2.2 Multiple Imputation Method (MI)

The multiple imputation method (MI) involves replacing each missing value with an average based on more than one simulated value [26-27]. The method applied in the present study differs from previous studies because the multiple imputation method involves repeating the SI method 10 times in two principal phases. The first phase consists of two steps. The first step involves imputing the missing values using single imputation method. The second step involves determining the estimated statistics through the use of the parameters obtained during the first imputation.

Similarly, the second phase consists of two steps, jointly referred to as a stage. The first step involves performing a second single imputation and estimating the parameters. The estimated parameter from the second imputation is averaged with the estimated parameter resulting from the first single imputation and applied when determining the estimated statistics based on the derived model. The averaging and application of the estimated parameter from the first and second imputation is referred to as Stage 1. Subsequent stages incorporate the estimated parameter of an additional single imputation, which is averaged alongside the estimated parameters included in the previous stage and applied in the derived model. The process is repeated nine times, resulting in the parameter estimates of ten separate single imputations being averaged and applied to the derived model at the ninth stage.

5 Performing the Imputation Method

The data at the Bukit Bendera target station are assumed to be missing data up to a certain percent for the purpose of testing the performance of the different methods on the model fitting. The missing hourly data for the majority of stations in Peninsular Malaysia is found to be less than or equal to than 10%. However, to investigate the consistency of the results with the model fitting results, the analysis examines three different percentages of missing data: 5%, 10% and 15%. The accuracy of the model is assessed using goodness of fit statistics involving the mean, variance, autocorrelation and probability of wet hours. The performance of the imputation methods for deriving the estimated statistics of the model are compared with the observed statistics of the complete data using the mean absolute error (**MAE**) criterion, which is as follows:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|1-\frac{X_{i-est}}{X_{i-obs}}\right| \tag{2}$$

where X_{i-est} stands for the value of the ith estimated statistics, X_{i-obs} is the corresponding observed value and m is the number of statistics evaluated.

6 Results and Discussion

Three methods of imputing missing hourly rainfall data are compared in terms of mean absolute error and the results are presented in Table 1. The SI method is found to dominate the ZS and MI methods for datasets where up to 10% of the data are missing. However, the MI method results demonstrate better levels of MAE from stage 1 to stage 9 for all missing data; and the estimates generally do not improve beyond stage 2, indicating that more than two repetitions of the same stage do not improve the estimated statistics. The results of 5% missing data are selected for comparing the estimated statistics, based upon the MBL model, with the corresponding observed statistics. Figures 4 to 7 demonstrate the observed and estimated statistics which are based on the mean, variance, lag-1 autocorrelation and probability of wet for both 1-hour rainfall data and 24-hour rainfall data. The results indicate that the single imputation method outperforms the other two methods as it produces a close agreement between the estimated and the observed statistics for all statistics with exception to the lag-1 autocorrelation. All of the methods fail to provide a precise estimate of the lag-1 autocorrelation.

7 Conclusion

Three different imputation methods are proposed in the present paper: the zero substitution (ZS) method; the single imputation (SI) method; and the multiple imputation (MI) method. The performance of the methods is examined in relation to the analysis of rainfall data from the Bukit Bendera rain gauge station. The investigation of the three imputation methods is assessed based on the mean absolute error (MAE) between the estimated and observed statistics derived from the Bartlett Lewis model. The methods are tested with different percentages of randomly generated missing rainfall values. The single imputation (SI) method is found to perform better in all months considered.

Table1 Comparison of imputation methods based on MAE with various percentages of missing values for the same month of different years

	November			September		
Method	5%	10%	15%	5%	10%	15%
Zero Substitution	4.12	15.15	13.30	4.38	8.33	17.99
Single imputation	**2.74**	12.25	**7.38**	**2.87**	**5.50**	13.29
Multiple imputation	4.85	**11.97**	9.09	8.64	5.33	**11.84**
Values of MAE for stages						
stage1	5.21	12.42	11.33	10.50	**5.33**	19.07
stage 2	**4.85**	12.20	9.23	**8.64**	8.73	18.48
stage 3	14.21	13.11	9.43	11.99	15.52	17.12
stage 4	13.25	12.84	10.06	12.30	17.44	18.79
stage 5	13.82	13.05	**9.09**	10.91	17.31	18.25
stage 6	14.56	11.99	9.73	11.36	18.79	18.37
stage 7	13.06	**11.97**	9.84	11.63	18.05	13.91
stage 8	13.33	12.15	12.19	11.92	17.69	11.89
stage 9	12.73	12.35	10.06	11.31	17.72	**11.84**

	June			March		
Method	5%	10%	15%	5%	10%	15%
Zero Substitution	4.13	12.76	19.72	13.62	13.34	8.26
Single imputation	**1.66**	**7.53**	12.41	**7.92**	**8.60**	22.87
Multiple imputation	4.83	8.10	**9.83**	16.66	9.46	**8.09**
Values of MAE for stages						
stage1	7.83	**8.10**	10.49	**16.66**	13.56	**8.09**
stage 2	7.52	8.98	10.24	41.34	10.49	10.89
stage 3	7.84	8.97	9.92	41.88	11.12	9.10
stage 4	5.95	9.00	**9.83**	42.84	12.16	9.53
stage 5	**4.83**	11.81	11.62	42.16	10.08	10.33
stage 6	5.45	10.68	11.66	41.69	**9.46**	8.89
stage 7	5.40	9.75	11.14	41.78	10.98	10.47
stage 8	9.48	10.61	12.18	41.07	11.61	10.85
stage 9	10.27	10.20	11.46	38.21	12.44	11.31

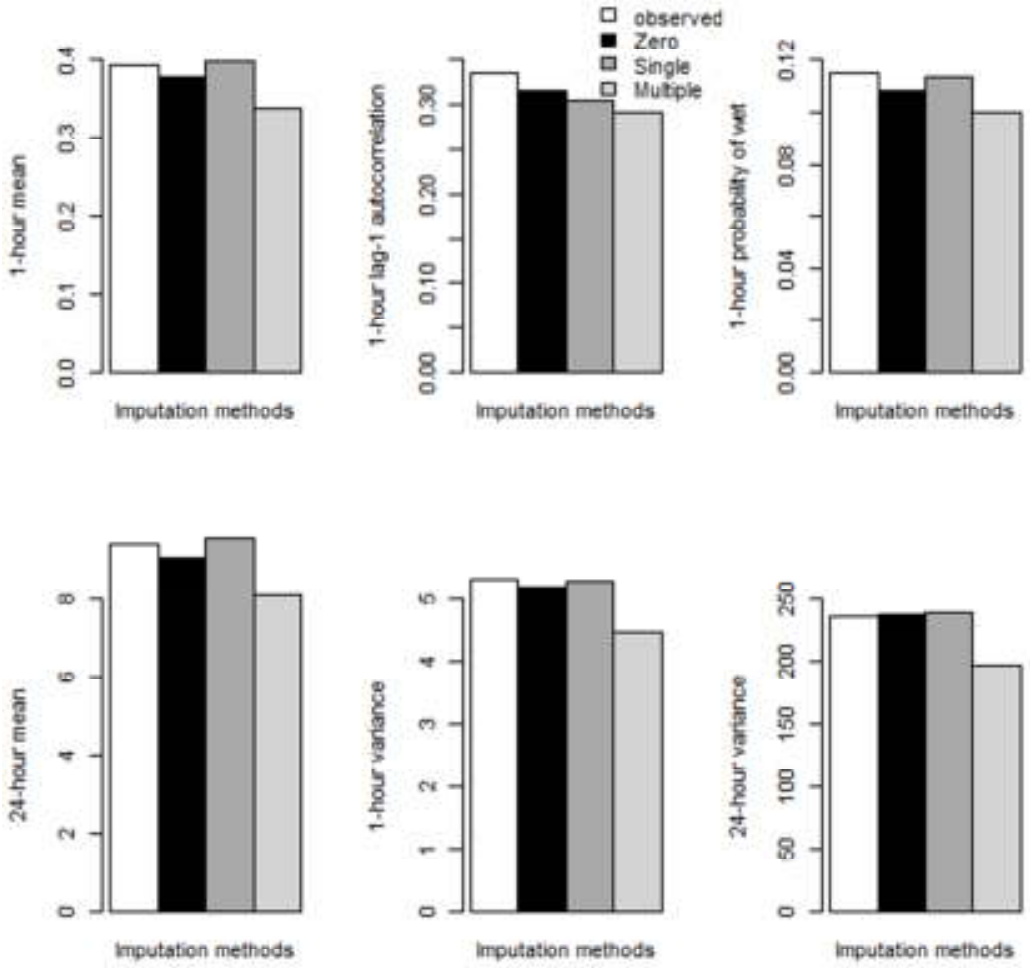

Figure 4 Comparison of the estimated statistics for three different imputation methods for November at Bukit Bendera station

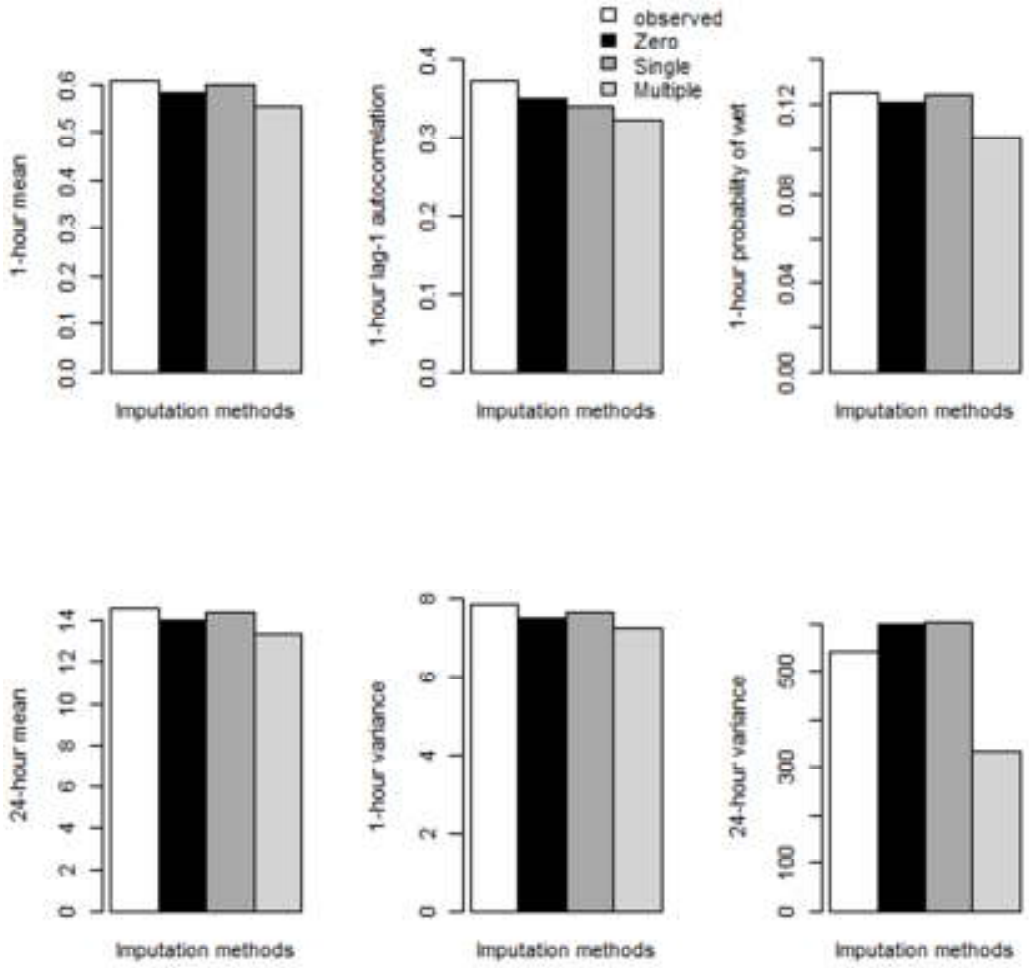

Figure 5 Comparison of the estimated statistics for three different imputation methods for September at Bukit Bendera station

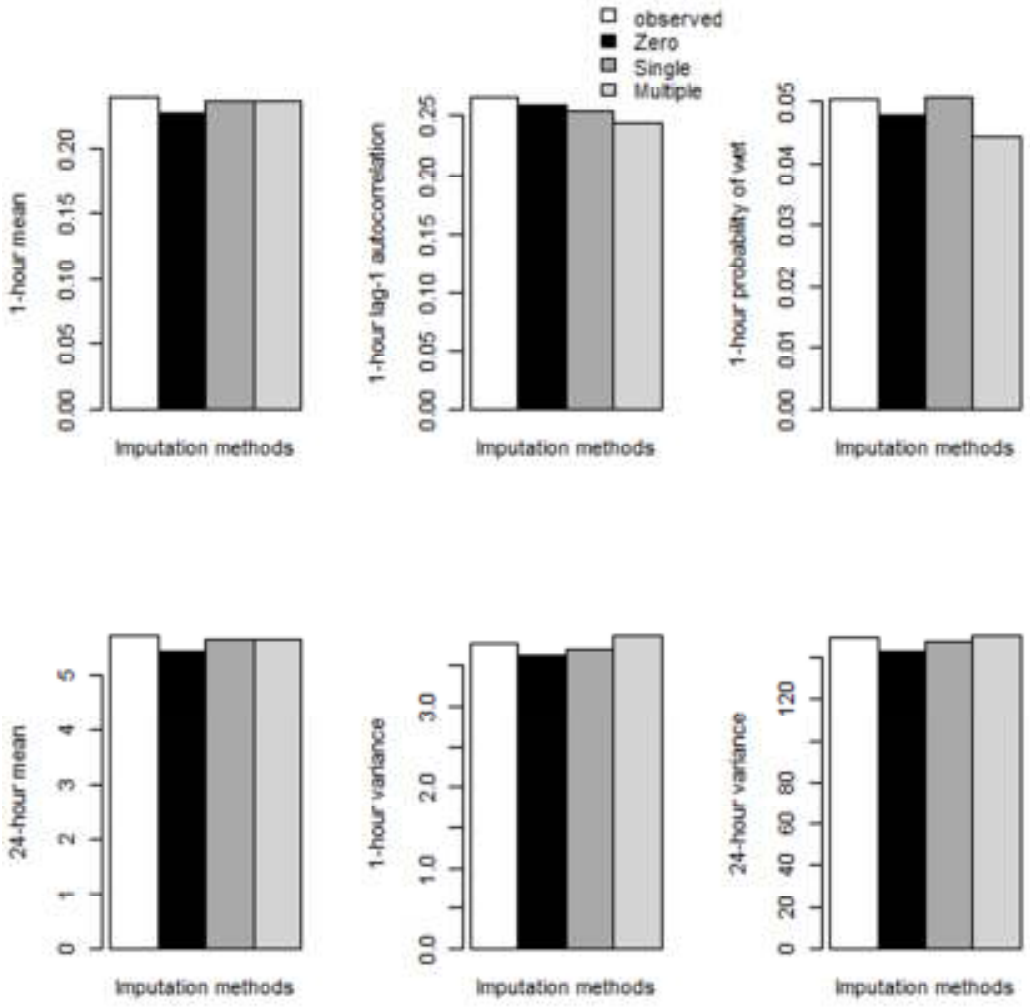

Figure 6 Comparison of the estimated statistics for three different imputation methods for June at Bukit Bendera station

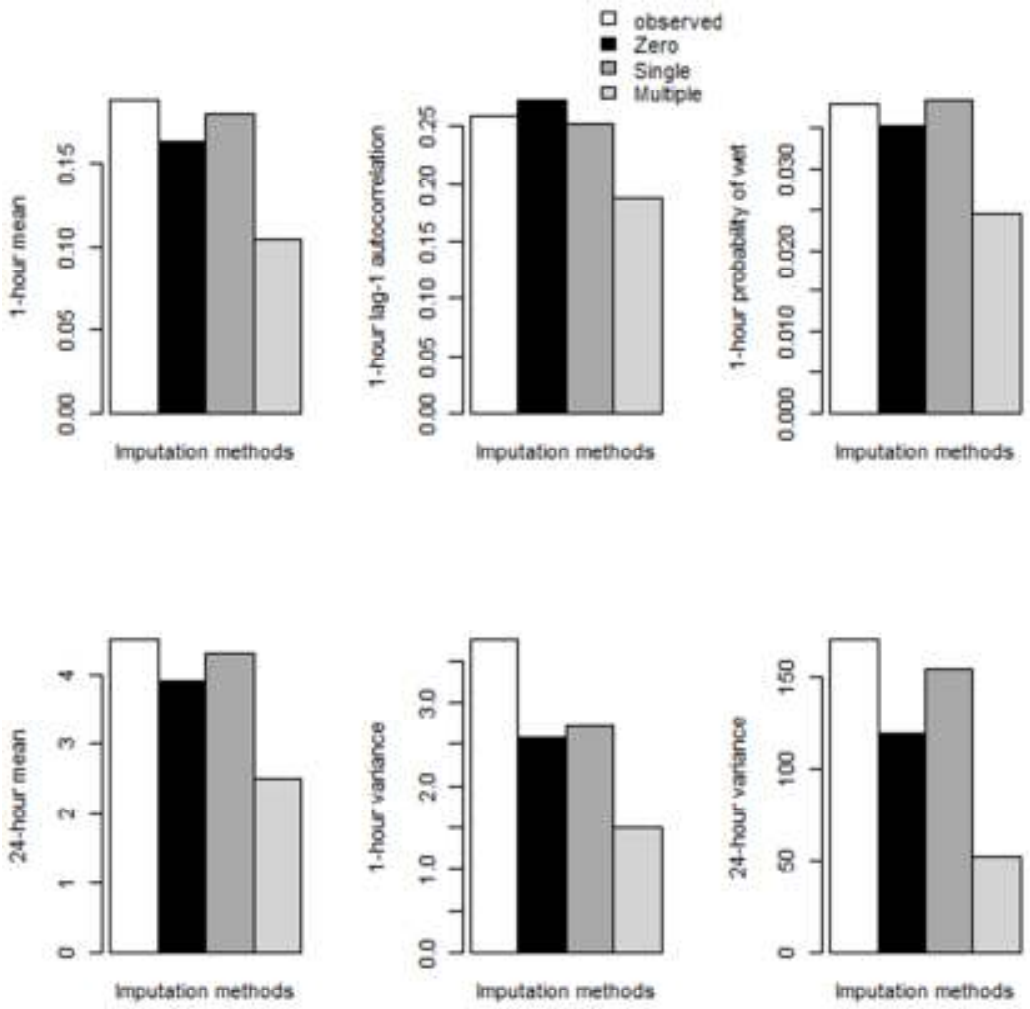

Figure 7 Comparison of the fitted statistics for three different imputation methods for March at Bukit Bendera station

Acknowledgements

The authors wish to thank Universiti Kebangsaan Malaysia for providing a partial support for this research.

References

[1] Malek, M.A., et al. Imputation of time series data via kohonen self organizing maps in the presence of missing data. *World Academy of Science, Engneering and Technology*. 2008. 17: 501-506.
[2] Paulhus, J.L.H. and Kohler, M.A. Interpolation of missing precipitation records. *Mon. Wea. Rev.* 1952. 80: 129-133.
[3] Teegavarapu, R.S.V. and Chandramouli, V. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *Journal of Hydrology.* 2005. 312: 191-206.
[4] Suhaila, J., Sayang, M.D. and Jemain, A.A. Revised spatial weighting methods for estimation of missing rainfall data. *Asia-Pacific Journal of Atmospheric Sciences.* 2008. 44(2): 93-104.
[5] Tang, W.Y., Kassim, A.H.M. and Abu Bakar, S.H. Comparative studies of various missing data treatment methods - Malaysian experience. *Atmospheric Research*. 1996. 42(1-4): 247-262.
[6] Shaowei, W., et al., *Estimating Wet-Pavement Exposure with Precipitation Data*, C.R. Project, Editor. California Department of Transportation (Caltrans) Sacramento. 2008.
[7] Verworn, A. and Haberlandt, U. Spatial interpolation of hourly rainfall – effect of additional information, variogram inference and storm properties. *Hydrology and Earth System Sciences.* 2011. 15: 569–584.
[8] Daly, C., Neilson, R.P. and Phillips, D.L. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor*. 1994. 33(2): 140-158.
[9] Rubin, D.B. Inference and missing data. *Biometrika.* 1976. 61: 581-592.
[10] Entekhabi, D., Rodriguez-Iturbe, I. and Eagleson, P. Probabilistic representation of the temporal rainfall process by a modified Neyman-Scott rectangular pulses model: parameter estimation and validation. *Water Resour. Res.* 1989. 25(2): 295-302.
[11] Rodriguez-Iturbe, I., Cox, D.R. and Isham, V. Some models for rainfall based on stochastic point processes. *Proc. R. Soc. Lond.* 1987a. A 410: 269-288.
[12] Rodriguez-Iturbe, I., Power, B.F.D. and Valdes, J.B. Rectangular pulses point process models for rainfall: analysis of empirical data. *J. Geophys. Res.* 1987b. 92: 9645-9656.
[13] Rodriguez-Iturbe, I., Cox, D.R. and Isham, V. A point process for rainfall: further developments. *Proceeding of the Royal Society London.* 1988. A417: 283-298.
[14] Cowpertwait, P. A generalized point process model for rainfall. *Proc. R. Soc. Lond.* 1994. A 447: 23-37.
[15] Cowpertwait, P. A Poisson-cluster model of rainfall: higher-order moments and extreme value. *Proceedings Royal Society London A.* 1998. 454: 885–898.
[16] Cowpertwait, P. A spatial–temporal point process model of rainfall for the Thames catchment, UK. *Journal of Hydrology.* 2006. 330: 586– 595.
[17] Onof, C., et al., Rainfall modelling using Poisson-cluster processes: a review of developments. *Stochastic Environmental Research and Risk Assessment*. 2000. 14(6): 384–411.
[18] Nelder, J.A. and Mead, R. A simplex method for function minimization. *The Computer Journal*. 1965. 7(4): 308–313.
[19] Spendley, W., Hext, G.R. and Himsworth, F.R. Sequential application of simplex designs in optimisation and evolutionary operation. *Technometrics.* 1962. ***4***: 441–461.
[20] Vanhaute, W.J., et al. Calibration of the modified Bartlett-Lewis model using global optimization techniques and alternative objective functions. *Hydrology and Earth System Sciences.* 2012. 16: 873–891.

[21] Reilly, M. Data analysis with hot deck multiple imputation. *The Statistician.* 1993. 42: 307-313.
[22] Little, R.J. and Rubin, D.B. *Statistical analysis with missing data*. New York: Wiley. 1987.
[23] Schafer, J.L. Dealing with missing data. *Res. Lett. Inf. Math. Sci*. 2002. 3: 153-160.
[24] Yusof, F., et al. Fitting the best-fit distribution for the hourly rainfall amount in the Wilayah Persekutuan. *Jurnal Teknologi.* 2007. 46C: 49-58.
[25] Shamsudin and Supiah. Probability distribution of rainfall depth at hourly time-scale. *International Conference on Environmental Science and Engineering,* Singapore. 2010.
[26] Rubin, D.B. *Multiple Imputation for Nonresponses in Surveys*. New York: J. Wiley & Sons. 1987.
[27] Schafer, J.L. Multiple imputation: A primer. *Statistical Methods in Medical Research.* 1999. 8: 3-15.