Comparison of Artificial Neural Network (ANN) and Other Imputation Methods in Estimating Missing Rainfall Data at Kuantan Station

Nur Afiqah Ahmad Norazizi[1] and Sayang Mohd Deni[1,2(✉)]

[1] Centre for Statistics and Decision Science Studies,
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
sayang@fskm.uitm.edu.my
[2] Advanced Analytic Engineering Center (AAEC),
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

Abstract. Daily rainfall data could be considered as one of the basic inputs in hydrological (e.g. streamflow, rainfall-runoff, recharge) and environmental (e.g. crop yield, drought risk) models as well as in assessing the water quality. In Malaysia, the number of rain gauge stations with complete records for a long duration is very scarce. The occurrence of missing values in rainfall data is mainly due to malfunctioning of equipment and severe environmental conditions. Thus, the estimation of rainfall is needed, whenever the missing data happened at the principal rainfall station. In this study, daily rainfall data from eight meteorological stations located in Pahang state are considered and Kuantan is selected as the target station. The main purposes of this study is to compare the performance of the imputation methods by using Artificial Neural Network method (ANN), Bootstrapping and Expectation Maximization Algorithm method and Multivariate Imputation by Chained Equations method (MICE). Missing rainfall data has been generated randomly for Kuantan station with 5%, 10% and 15% of missingness. The three methods are compared based on Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Coefficient of Determination (R2). The findings concluded that Artificial Neural Network (ANN) is found to be the best imputation method for this study, followed by Multiple Imputation by Chained Equation (MICE) and Bootstrapping and Expectation Maximization Algorithm method.

Keywords: Daily rainfall data · Artificial Neural Network · Bootstrapping and Expectation Maximization Algorithm · Multivariate Imputation by Chained Equations · Imputation method · Missing data

1 Introduction

Missing data is one of the common problems arise in the process of data collection. Incomplete data occurs for a variety of reasons, such as, interruption of experiments, equipment failure, measurement limitation, attrition in longitudinal studies, censoring,

M. W. Berry et al. (Eds.): SCDS 2019, CCIS 1100, pp. 298–306, 2019.
https://doi.org/10.1007/978-981-15-0399-3_24

usage of new instruments, changing methods of record keeping, lost of records, and non response to questionnaire items [1, 2, 3, 5, 8, 11, 14, 15, 18, 21–28].

There are various types of data driven models including Artificial Neural Network (ANN) which could be used for the implementation of weather forecast. The ANN provides good approximation due to the capability of the network, dynamic and works well with non-stationary data. ANN is a popular method for many hydrological data analyses as demonstrated by many researchers [6, 7, 9, 12–15, 19]. The ANN has been shown to be one of the best methods for missing data prediction at par with fuzzy logic as a fuzzy rule-based approach. The estimation of missing rainfall data using ANN were compared with the results obtained using regression and other simpler techniques such as arithmetic and inverse distance method. Based on some previous studies, the ANN was chosen to be compared with other techniques due to its adequacy and reliable in predicting missing rainfall at particular gauge stations. Meanwhile, other methods requires much longer duration of time to estimate the missing values and the process involved much more complicated if compared with ANN [3, 21].

Thus, the aim of this study is to estimate the missing rainfall data by using Artificial Neural Network (ANN), Bootstrapping and Expectation Maximization Algorithm and Multivariate Imputation by Chained Equations (MICE). These methods were chosen due to the successfulness and the capability in handling the missing value problems as mentioned by some previous researchers in the field of study [2, 4, 6, 7, 9, 12, 15, 18, 19]. In evaluating those methods, three different level of missing values such as 5%, 10% and 15% will be considered. The performance of these methods is assessed using the Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Correlation of Determination (R^2). In addition, the results of this study can provide knowledge to the researchers on several alternatives techniques for hydrological data in Malaysia and worldwide on which technique would give the best performance. Research and development in this study is conducted on continuous basis, thus there might be new findings on the issue of environmental as well.

2 Materials and Methods

2.1 Data Description

In this study, daily rainfall data were obtained from the Malaysian Meteorological Department and Drainage and Irrigation Department of Malaysia. The secondary data consists of daily rainfall amount (mm) for the period of 1975 up to 2017. Kuantan station which is located in the state of Pahang is chosen due to heavy seasonal rainfall and winds that affected most parts of Peninsular Malaysia during December 2014. The rains caused severe flooding in the East Coast region i.e. Terengganu, Pahang, and Kelantan states [17]. There were eight meteorological stations located in Pahang were selected such as Sekolah Menengah Ahmad Pekan, Felda Bukit Tajau, Felda Kampung New Zealand, Felda Sungai Pancing Selatan, Pusat Pertanian Tanaman Kampung Awah, Pusat Pertanian Bukit Goh and Mardi Sungai Baging as well as Kuantan station was chosen as the target station (Table 1).

Table 1. The geographical coordinates of the selected rainfall stations

Station	Latitude (N)	Longitude (E)
Felda Bukit Tajau	3.58°	102.73°
Felda Kampung New Zealand	3.64°	102.86°
Felda Sungai Pancing Selatan	3.80°	103.17°
Kuantan	3.76°	103.21°
Mardi Sungai Baging	4.08°	103.39°
Pusat Pertanian Tanaman Kampung Awah	3.48°	102.52°
Pusat Pertanian Bukit Goh	3.88°	103.26°
Sekolah Menengah Kebangsaan Ahmad, Pekan	3.49°	103.40°

2.2 Missing Data Imputation Analysis by Using Artificial Neural Network

In the process of imputation of missing values, the data was trained using ANN with three different learning algorithms, namely, conjugate gradient Fletcher–Reeves update (CGF), Broyden–Fletcher–Goldforb–Shanno (BFG) and Levenberg–Marquardt (LM). Three different units of analysis such as, daily, 10-day and monthly rainfall values, were used for evaluating the prediction ability of ANN. The output of the network was identified as the amount of precipitation at station X, px(t), with the inputs which was determined by the amount of the neighboring stations within the same duration in the period of time. The equation of the model assessment can be expressed as:

$$p_x(t) = f[p_1(t), p_2(t), p_3(t), p_4(t), p_5(t), p_6(t), p_7(t)] \tag{1}$$

where p1, p2, p3, p4, p5, p6 and p7 are denoted as the amount of rainfall at seven neighboring stations except Kuantan station. The ANN was trained and simulated using R Programming with the *neuralnet* package.

2.3 Missing Data Imputation Method by Using Bootstrapping and Expectation Maximization Algorithm

One of the advantages of Bootstrapping and Expectation Maximization Algorithm in Amelia package of R Programming is that, the combination of speed and the ease-of-use of algorithm with the power of multiple imputations will take into consideration.

The imputation model in Amelia package assumes that the complete data (that is, both observed and unobserved) are multivariate normal. It is denoted as the ($n \times k$) dataset with D is defined as the observed part, D_{obs} and unobserved part, D_{mis}, where the assumption is given as the following equation:

$$D \sim N_k(\mu, \Sigma) \tag{2}$$

The state, D is defined as multivariate normal distribution with mean vector μ and covariance matrix Σ. The multivariate normal distribution is often to give a crude approximation to the true distribution of the data. Moreover, there is evidence to show

that this model works as well as the other models which are more complicated and contained the mixed data. It has been reported by some previous researchers that multivariate normal model can provide valid estimate even though the assumptions is violated. This may be due to the large sample size and small percentage of missing values occurred in the dataset [3, 21]. Furthermore, transformations of many types of variables can often make this normality assumption more plausible.

2.4 Missing Data Imputation Analysis by Using Multivariate Imputation by Chained Equation

Multiple Imputation by Chained Equations (MICE) is a practical approach to generate missing rainfall values based on a set of imputation models. MICE is also known as fully conditional specification and sequential regression multivariate imputation. The three stages of MICE are described as below:

i. Generating Multiple Imputed Data Sets
ii. Analyzing Multiple Imputed Data Sets
iii. Combining Estimates from Multiply Imputed Data Sets

The m estimates are combined into an overall estimate and variance–covariance matrix using Rubin's rules, which are based on asymptotic theory in a Bayesian framework,

$$\hat{\theta}_j = \frac{1}{m}\sum\nolimits_{j=1}^{m} \hat{\theta}_j \tag{3}$$

2.5 Performance Measure Criteria

Three performance indicators which are Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Coefficient of Determination (R2) will be used to assess the imputation methods.

$$MAE = \frac{1}{n}\sum\nolimits_{i=1}^{n} |f_i - y_i| = \frac{1}{n}\sum\nolimits_{i=1}^{b} |e_i| \tag{4}$$

where f_i is the prediction and y_i is the true value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(X_{obs,i} - X_{model,i}\right)^2}{n}} \tag{5}$$

Where X_{obs} is the observed values and X_{model} is the modeled values, at time or location i.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum y_i^2} = \frac{ESS}{TSS} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} \tag{6}$$

3 Results and Discussion

Referring to Table 2, the lowest mean of rainfall amount of 5.3 mm is recorded at Pusat Pertanian Tanaman Kampung Awah. Meanwhile, the highest mean of rainfall amount of 8 mm is observed at Felda Sungai Pancing Selatan. For the period of from 1975 up to 2017, it is observed that Kuantan station is having a complete of records with no missing values. Meanwhile, 9.4% of missing values is found from Felda Bukit Tajau station which is the most missing values recorded compared to the rest of neighboring stations.

Table 2. Descriptive statistics for the target station and the neighboring stations

Station	Minimum value (mm)	Maximum value (mm)	Mean (mm)	Percentage of missing (%)
Felda B. Tajau	0.0	176.9	5.6	9.4
Felda Kg. New Zealand	0.0	184.5	5.8	3.2
Felda Sg. Pancing Selatan	0.0	280.5	8.0	1.6
Kuantan	0.0	527.5	7.9	0.0
Mardi Sg. Baging	0.0	541.5	7.7	0.7
P.P. Tanaman Kg. Awah	0.0	176.8	5.3	1.3
P.P.B. Goh	0.0	407.6	7.7	0.8
SMK Ahmad, Pekan	0.0	402.4	7.7	3.6

The visualization of the computed Artificial Neural Network is shown below. The model has 3 hidden units including the neurons and layers. The black lines showed connections with the weights of the values which will be used to impute the missing values in the target station. In addition, the blue line demonstrated the bias item which is generated by the neural network estimation. The figure showed the results for the 5% generated missing values for Kuantan station (Fig. 1).

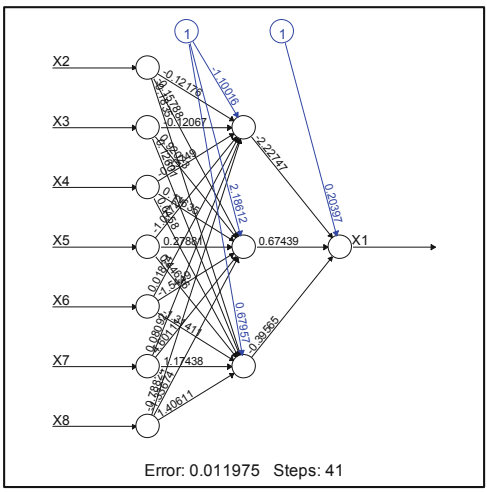

Fig. 1. Visualization by using Artificial Neural Network (Color figure online)

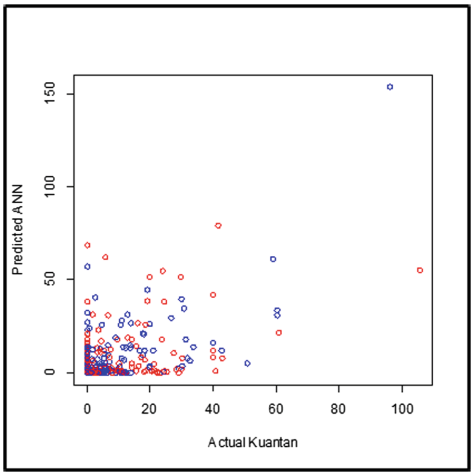

Fig. 2. Predicted values by using Artificial Neural Network versus actual values in Kuantan station (5%) (Color figure online)

Figure 2 shows the graph of the comparison of the actual data and the predicted values imputed by using Artificial Neural Network (ANN) for 5% of missing values at Kuantan station. For Fig. 2 up to Fig. 4, the blue and red dots are denoted as the actual values and the predicted values, respectively.

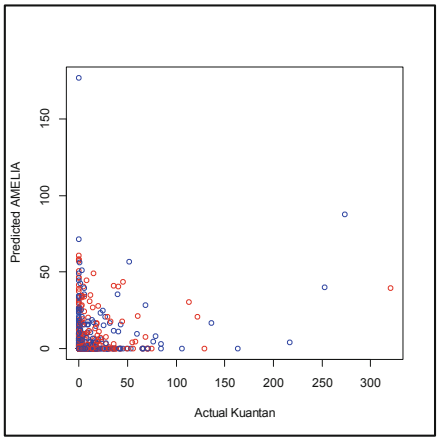

Fig. 3. Predicted values by using Bootstrapping and Expectation Maximization Algorithm versus actual values in Kuantan station (5%) (Color figure online)

Figure 3 showed the plot of comparison of the actual data and the predicted values imputed by using Expectation Maximization Algorithm estimation process for 5% of missing values at Kuantan station.

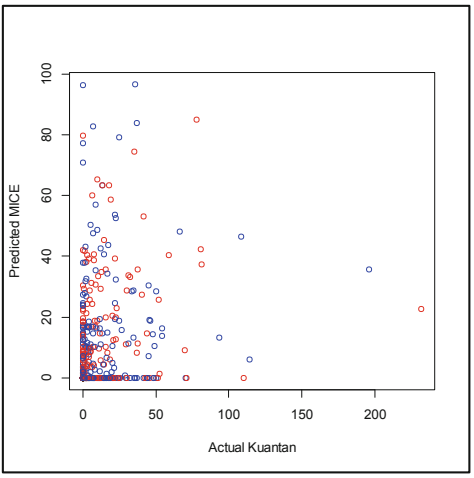

Fig. 4. Predicted values by using Multivariate Imputations by Chained Equation versus actual values in Kuantan station (5%) (Color figure online)

Figure 4 demonstrated the plot of comparison of the actual data and the predicted values imputed by using Multivariate Imputations by Chained Equation process for 5% of missing values at Kuantan station.

Table 3. Performance measures for Mean Absolute Error, Root Mean Squared Error and Coefficient of Determination

R package	MAE			RMSE			R^2		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
NEURALNET	6.460	7.531	7.852	16.798	15.832	14.176	0.978	0.943	0.972
AMELIA	9.111	9.387	9.444	23.888	22.196	22.165	0.755	0.731	0.712
MICE	8.604	9.242	9.091	19.520	18.283	19.511	0.780	0.877	0.908

In order to determine the best imputation method, the lowest value of MAE and RMSE will be chosen as well as the highest value for the R square. Based on the results showed in Table 3, it could be concluded that the Artificial Neural Network (ANN) is observed to be the best imputation method followed by Multiple Imputation by Chained Equation (MICE) and Bootstrapping and Expectation Maximization Algorithm method. It is observed that the result produced by MAE and RMSE are consistently showed the lowest, and the highest value of R square for NEURALNET and followed by MICE and finally by AMELIA for the three level of missing values. Thus, it could be concluded that based on the data used in the study, the Artificial Neural Network (ANN) is found to be the best imputation method in generating missing rainfall data at Kuantan station.

4 Conclusion

The aim of this research is to compare three imputation methods in imputing the missing values at Kuantan station due to the completeness in the data set for the period of 1975 up to 2017. In evaluating the three imputations methods, missing data were created at three different levels, 5%, 10% and 15%. In addition, missing data at Kuantan station has been generated using Missing at Random (MAR) assumption. The predicted imputation results for each method were compared with the actual data at this station.

The performance for each method was evaluated using the three performance measures such as MAE, RMSE and R Squared. According to some previous studies, the most popular method for imputation of rainfall data is not necessarily to be most efficient. However, the results shown that the Artificial Neural Network (ANN) by using *neuralnet* package in R demonstrated the best method estimation for imputing the missing rainfall data compared to the two other methods. Meanwhile, Bootstrapping and Expectation Maximization Algorithm in Amelia package and Multivariate Imputation by Chained Equation (MICE) in MICE package of R Programming are found rarely been used for the estimation of missing rainfall data. However, these two methods and R programming packages which are Amelia and MICE, are both widely used in the imputation of missing data only in other extent aside from missing rainfall data.

Acknowledgement. The authors wish to thank Malaysian Meteorological Department for the data and sponsorship from Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM). The authors are also indebted to the staff of the Drainage and Irrigation Department for providing the daily rainfall data for this study. They also acknowledge their sincere appreciation to the reviewers for their valuable suggestion and remarks in order to improve the manuscript. This research will not complete without the sponsorship from Ministry of Higher Education (600-RMI/TRGS DIS 5/3 (1/2015)).

References

1. Abuelgasim, A.A., Gopal, S., Strahler, A.H.: Forward and inverse modelling of canopy directional reflectance using a neural network. Int. J. Remote Sens. **19**(3), 453–471 (1998)
2. Amer, S.R.: Neural network imputation: a new fashion or a good tool. Unpublished Ph.D. thesis (2004)
3. Demirtas, H., Freels, S.A., Yucel, R.M.: Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. JSCS **78**(1), 69–84 (2008)
4. Fausett, L.: Fundamentals of Neural Networks: Architectures, Algorithms, and Applications (No. 006.3). Prentice-Hall (1994)
5. Kamaruzaman, I.F., Zin, W.Z.W., Ariff, N.M.: A comparison of method for treating missing daily rainfall data in Peninsular Malaysia. Malays. J. Fundam. Appl. Sci. (4–1), 375–380 (2017)

6. Khan, I.Y., Zope, P.H., Suralkar, S.R.: Importance of artificial neural network in medical diagnosis disease like acute nephritis disease and heart disease. Int. J. Eng. Sci. Innov. Technol. (IJESIT) **2**(2), 210–217 (2013)
7. Karunanithi, N., Grenney, W.J., Whitley, D., Bovee, K.: Neural networks for river flow prediction. J. Comput. Civ. Eng. **8**(2), 201–220 (1994)
8. Shaadan, N., Deni, S.M., Jemain, A.A.: Application of functional data analysis for the treatment of missing air quality data. Sains Malays. **44**(10), 1531–1540 (2015)
9. Kuligowski, R.J., Barros, A.P.: Using artificial neural networks to estimate missing rainfall data. J. Am. Water Resour. Assoc. **34**(6), 1437–1447 (1998)
10. Le Barbé, L., Lebel, T., Tapsoba, D.: Rain fall variability in West Africa during the years 1950–1990. J. Clim. **15**(2), 187–202 (2002)
11. Leung, H., Haykin, S.: Detection and estimation using an adaptive rational function filter. IEEE Trans. Signal Process. **42**(12), 3366–3376 (1994)
12. Livingstone, D.J., Manallack, D.T., Tetko, I.V.: Data modeling with neural networks–an answer to the maiden's prayer? J. Comp. Aid. Mol. Des. **11**, 135–142 (1996)
13. Nasr, M., Zahran, H.F.: Using of pH as a tool to predict salinity of groundwater for irrigation purpose using artificial neural network. Egypt. J. Aquat. Res. **40**(2), 111–115 (2014)
14. Rahman, N.A., Deni, S.M., Ramli, N.M.: Generalized linear model for estimation of missing daily rainfall data. AIP Conf. Proc. **1830**(1), 080019 (2017)
15. Paulhus, J.L., Kohler, M.A.: Interpolation of missing precipitation records. Mon. Weather Rev. **80**(8), 129–133 (1952)
16. Ratnayake, U., Herath, S.: Changing rainfall and its impact on landslides in Sri Lanka. J. Mt. Sci. **2**(3), 218–224 (2005)
17. ReliefWeb (2014). Malaysia: Seasonal Floods 2014 - Information Bulletin n° 1. https://reliefweb.int/report/malaysia/malaysia-seasonal-floods-2014-aaainformation-bulletin-n-1. Accessed 4 Nov 2017
18. Royston, P., White, I.R.: Multiple imputation by chained equations (MICE): implementation in Stata. J. Stat. Softw. **45**(4), 1–20 (2011)
19. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533 (1986)
20. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)
21. Schaming, D., et al.: Easy methods for the electropolymerization of porphyrins based on the oxidation of the macrocycles. Electrochim. Acta **56**(28), 10454–10463 (2011)
22. Schafer, J.L.: Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC, London (1997)
23. Burhanuddin, S.N.Z.A., Deni, S.M., Ramli, N.M.: Normal ratio in multiple imputation based on Bootstrapped sample for rainfall data with missingness. Int. J. Geomate **13**(36), 131–137 (2017)
24. Suhaila, J., Jemain, A.A., Hamdan, M.F., Zin, W.Z.W.: Comparing rainfall patterns between regions in Peninsular Malaysia via a functional data analysis technique. J. Hydrol. **411**(3), 197–206 (2011)
25. Suhaila, J., Sayang, M.D., Jemain, A.A.: Revised spatial weighting methods for estimation of missing rainfall data. Asia Pac. J. Atmos. Sci. **44**(2), 93–104 (2008)
26. Tang, W.Y., Kassim, A.H.M., Abu Bakar, S.H.: Comparative studies of various missing data treatment methods-Malaysian experience. Atmos. Res. **42**(1–4), 247–262 (1996)
27. Von Davier, M.: Imputing proficiency data under planned missingness in population models. In: Handbook of International Large-Scale Assessment. Background, Technical Issues, and Methods of Data Analysis, pp. 175–202 (2014)
28. Young, K.C.: A three-way model for interpolating for monthly precipitation values. Mon. Weather Rev. **120**(11), 2561–2569 (1992)