MJFAS MALAYSIAN JOURNAL OF FUNDAMENTAL AND APPLIED SCIENCES
PRINT ISSN: 2289-5981 | ONLINE ISSN: 2289-599X

RESEARCH ARTICLE

Estimation of rainfall and stream flow missing data for Terengganu, Malaysia by using interpolation technique methods

Wan Norliyana Wan Ismail*, Wan Zawiah Wan Zin@Wan Ibrahim

School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia

* *Corresponding author: wanna2573@gmail.com*

Article history
Received 18 Feb 2017
Accepted 25 August 2017

Abstract

Missing data is a serious problem in many climatological time series. Daily rainfall and stream flow datasets with no missing values are required for efficient estimation for application purposes. In order to estimate any missing observations in data, interpolation techniques are often used. This study focuses on comparing a few selected methods in the estimation of missing rainfall and stream flow data. The interpolation techniques studied were the Arithmetic Average (AA) method, Normal Ratio (NR) method, Inverse Distance (ID) method and Coefficient of Correlation (CC) method. However, in the case when there is no information from neighboring stations, the mean on the same day and month but at different years is taken as estimation of the missing value on that particular date. Twenty years of daily rainfall and stream flow data at 12 stations located at Terengganu were used for this study. In testing to verify which method is the best in evaluating missing values at the target station using information from the nearby stations (in the radius range of 10 km to 50 km), several percentages of missing values were considered. The validation of the best estimation methods is done based on the estimation error; with tests such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Correlation Coefficient (R) tests.

Keywords: Missing data, interpolation method, rainfall and stream flow data

INTRODUCTION

Rainfall is needed as a source of fresh water, which is essential for the survival of humans, plants and animals as well as stream flow availability. Rainfall and stream flow plays a significant role in hydrological, agricultural models and in assessing water quality. Studying about rainfall is important in order to identify the rainfall characteristics, occurrence of spatial and temporal variability, forecasting extreme rainfall events and hence, the problems such as floods, droughts and landslides may be resolve. Meanwhile, stream flow cycle is the section where rainfall occurs and results in flow. Floods also can happen when the volume of water exceeds the capacity of the river.

Rainfall and stream flow may contain missing values which attributed to various reasons such as bad weather, instrumental failures or human error during data entry (De Silva *et al.*, 2007; Suhaila *et al.*, 2008). Estimation of missing values becomes first priority in the data preparation process. In order to have completeness of data, missing data treatment is a necessary procedure to perform statistical analysis.

Various methods have been done to estimate missing rainfall and stream flow data. (Hasan and Croke, 2013) discuss a probabilistic approach and also interpolation method for matching the data point in daily rainfall series. Poisson-gamma (PG) models were compared with inverse distance interpolation method. However, PG models do not capture well the large rainfall events but still performs better than interpolation method. Artificial neural networks (ANNs) and adaptive neuro-fuzzy inference system (ANFIS) methods were used by (Dastorani *et al.*, 2010) to predict the missing flow data. Two common methods such as the normal ratio (NR) method and the correlation method were also employed, and by comparing those four methods, ANN was found as an efficient method for estimation of missing data.

Spatial interpolation methods refer to the process of estimating the unknown data values for a point using the known data values from nearby stations (Burrough and McDonnell, 1998). (Paulhus and Kohler, 1952) proposed normal ratio (NR) method of spatial interpolation which is based on the ratio means of data between the target station and the neighbouring station. This method is the most common interpolation used in estimating missing rainfall data (Chow et al, 1988).

(Suhaila *et al.*, 2008) explore the inverse distance (ID) weighting method as one of the simpler methods which the assumption of the rainfall values at the target station could be in influenced most by the nearest stations and less by the more distant stations. Correlation coefficient (CC) is another spatial interpolation method proposed by (Teegavarapu and Chandramouli, 2005) to estimate the missing rainfall data at 20 rain gauge station which is by replacing the weighting value of the ID method with the CC method.

Another simpler method to estimate missing data is arithmetic average (AA) method which evaluates the mean annual rainfall amount at the target station as well as the nearby stations. However, there are still missing values detected in the data which could not be complete although after using spatial interpolation methods. In order to predict the missing target data, (Ibrahim and Wibowo, 2014) apply temporal interpolation method which use the mean of corresponding months to replace the missing values of rainfall and water level.

The main purposes of this study are to evaluate missing rainfall and stream flow data using several interpolation methods which are aarithmetic average (AA) method, normal ratio (NR) method, inverse ddistance (ID) method, and coefficient of correlation (CC) method. However, if the data are still missing and the information from the neighbouring stations cannot be used because of the lacking of data, the mean on the same day and month but at different year is taken as the estimation of the missing values on that particular date. To evaluate the missing values at the target station using the information from neighbouring stations, the analysis are divided into four different percentages namely 5%, 10%, 15% and 20% in order to represent various cases of missing data. Besides, the performance of those methods are compared based on Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Correlation Coefficient (R) tests.

Area of study

Terengganu is located between 5°19' North latitude and 103°8' East longitude. Terengganu have eight districts which Kuala Terengganu is the capital city of Terengganu. Terengganu experiences dry season from May until June and experiences heavy rainfall during the northeast monsoon in November and December. In this study, 12 stations located at Terengganu have been used. Six stations are considered as the target station for rainfall data and six other target stations are for stream flow data. Each target station involve two neighbouring stations in the radius of 10 km to 50 km, refer to Fig. 1. The data consist of daily rainfall and stream flow amount from 1995 to 2014 (20 years). Both data were obtained from Department of Irrigation and Drainage, Ampang, Selangor. Those years are chosen based on the completeness available period of the data. The list of stations are provided in Table 1.

Fig. 1 The location of the target station and the neighboring stations in Terengganu.

Table 1 Description of the six rainfall stations and six stream flow stations in Terengganu within 50 km used as neighboring stations.

Station Number	Station Name	Latitude	Longitude	Neighboring Station	Euclidean Distance (km)
		RAINFALL			
TRa (4131001)	Kg. Ban Ho	4.133	103.175	TRb, TRc	15.51, 28.87
TRb (4232002)	Jambatan Air Putih	4.271	103.199	TRa, TRc	15.51, 13.84
TRc (4332001)	Jambatan Tebak	4.378	103.263	TRa, TRb	28.87, 13.84
TRd (4529001)	Rumah Pam Paya Kempian, Pasir Raja	4.568	103.979	TRe, TRf	17.67, 22.89
TRe (4730002)	Jambatan Jerangau	4.735	103.088	TRd, TRf	17.67, 19.10
TRf (4832011)	Kg. Menerong	4.843	103.204	TRd, TRe	22.89, 19.10
		STREAM FLOW			
TSa (4131453)	Sg. Cherul, Ban Ho	4.133	103.175	TSb, TSc	15.51, 28.86
TSb (4232401)	Sg. Kemaman, Jam. Air Putih	4.271	103.199	TSa, TSc	15.51, 13.84
TSc (4332401)	Sg. Tebak, Jam. Tebak	4.378	103.263	TSa, TSb	28.86, 13.84
TSd (4832441)	Sg. Dungun, Jam. Jerangau	4.843	103.204	TSe, TSf	18.97, 37.14
TSe (4930401)	Sg. Berang, Menerong	4.939	103.062	TSd, TSf	18.97, 22.16
TSf (5130432)	Sg. Terengganu, Kg. Tanggol	5.138	103.046	TSd, TSe	37.14, 22.16

RESEARCH METHODOLOGY

This section is divided into two main subsections. Methods for estimating missing data will be discussed in the first subsection. The analysis involved a target and some selected neighbouring stations. Meanwhile, assessing the performance of the methods used will be described in the second subsection. In first section, the target station has a complete set of data. Then, for the purpose of testing the estimation methods, data at the target station are assumed to be missing. By using the interpolation methods, the missing rainfall and stream flow data in target station are compared with the actual records. However, if there are still missing data after using interpolation methods, the mean on the same day and month but at different years will be used. Missing value of target data is replaced by the mean of the non-missing data from 1995 until 2014. Other than that, in order to have better estimation results, different percentages namely 5%, 10%, 15% and 20% are compared to represent various cases for missing data.

Interpolation methods

(i) Arithmetic Average Method

The arithmetic average (AA) method is the simplest method which is commonly used to fill the missing meteorological and hydrological data. The missing data of rainfall and stream flow are obtained by the average of selected nearby stations around the target station or the date on the same day with different years. The estimated missing value is given by

$$p_t = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

where p_t is the estimated value of the missing data at the t target station/date, x_i is the observed data at ith nearby stations or the date of the same date with different years and n is the number of nearby stations or number of years.

(ii) Normal Ratio Method

Normal ratio (NR) method is weighted based on the ratio mean of the available data between the target station and the *i*th neighboring station. This method is used if any neighboring stations have the normal annual rainfall and stream flow data which exceeded more than 10% of the considered station (Silva et al, 2007). The estimated missing value is given by

$$p_t = \frac{1}{n}\sum_{i=1}^{n}\frac{N_t}{N_i}x_i \tag{2}$$

where N_t is the annual rainfall and stream flow amount at the target station and N_i is the annual rainfall and stream flow amount at the *i*th nearby station.

(iii) Inverse Distance Method

Inverse distance (ID) method is the most commonly used for estimation of missing data. In this method, it is based on the distance between target station and nearby station. The closer stations are better correlated with the target station compared to further stations. The estimated missing value is given by

$$p_t = \frac{\sum_{i=1}^{n} x_i / d_{it}}{\sum_{i=1}^{n} 1 / d_{it}} \tag{3}$$

where d_{it} is the distance between target station and the *i*th nearby station.

(iv) Coefficient of Correlation Method

Coefficient of correlation (CC) method is influenced by the success of the ID method. This method is used by replacing the distance with the correlation coefficient as the weighting value. The estimated missing value is given by

$$p_t = \frac{\sum_{i=1}^{n} x_i r_{it}}{\sum_{i=1}^{n} r_{it}} \tag{4}$$

where r_{it} is the correlation coefficient of daily time series data between the target station and the *i*th nearby station.

Performance of the estimation methods

In this study, three performance criteria are used. The root mean square errors (RMSE), the mean absolute errors (MAE) and the correlation coefficient (R) statistics are calculated to evaluate spatial interpolation methods. The error measures the difference between the estimation values and their corresponding observed values. RMSE and MAE which indicate lower values will give better performances. Meanwhile, correlation coefficient indicates the strength of the relationship between observations and estimates which the higher positive coefficients estimate the best results.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{x}_i - x_i| \tag{6}$$

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\hat{x}_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(\hat{x}_i - \bar{x})^2}} \tag{7}$$

where x_i is the observed rainfall and stream flow at nearby station, $\hat{x}_i$ is the estimated value and is the number of nearby station.

RESULTS AND DISCUSSION

In this section, we will discuss briefly the results of the analysis. All four interpolation methods have been tested on four different percentages at 5%, 10%, 15% and 20%. The results of the overall performance of the methods are reported in Table 2 and Table 3. Table 2 shows the comparison of estimation methods for rainfall data. The ID method is found to be the best methods for TRb, TRc and TRf stations. For TRa and TRe stations, both stations recorded NR as the best method. It is also shown that CC method is the second best method for all stations which gave the minimum RMSE

Table 2 Comparison of estimation methods based on RMSE, MAE and R with four different percentages of missing values for rainfall data.

Station	Methods	RMSE				MAE				R			
		5%	10%	15%	20%	5%	10%	15%	20%	5%	10%	15%	20%
TRa	AA	15.803	16.831	15.805	16.793	7.155	7.290	7.037	7.292	1.593	2.779	3.196	4.039
	ID	16.392	17.158	16.003	16.702	7.369	7.550	7.188	7.323	1.637	2.867	3.309	4.202
	NR	15.979	16.650	15.761	16.497	7.241	7.220	7.019	7.182	1.592	2.775	3.196	4.047
	CC	15.806	16.834	15.805	16.775	7.159	7.290	7.037	7.287	1.592	2.778	3.196	4.043
TRb	AA	14.785	14.076	12.895	13.478	6.743	6.581	6.096	6.241	1.558	2.883	3.276	4.209
	ID	14.760	13.943	12.765	13.411	6.704	6.514	6.042	6.204	1.547	2.872	3.258	4.188
	NR	14.738	14.361	12.923	13.470	6.713	6.788	6.119	6.227	1.557	2.876	3.275	4.213
	CC	14.769	13.991	12.806	13.431	6.728	6.541	6.061	6.219	1.555	2.877	3.265	4.198
TRc	AA	15.028	14.786	13.861	15.555	6.794	6.803	6.591	6.915	1.495	2.729	3.099	3.954
	ID	15.615	14.680	13.607	15.163	6.906	6.801	6.453	6.733	1.477	2.759	3.119	4.009
	NR	15.023	14.908	13.868	15.567	6.742	6.714	6.590	7.015	1.496	2.731	3.099	3.947
	CC	15.016	14.685	13.731	15.406	6.787	6.781	6.551	6.869	1.495	2.732	3.102	3.963
TRd	AA	19.663	18.166	17.080	16.562	8.237	8.149	8.190	8.094	1.920	3.103	3.756	4.534
	ID	19.500	18.517	17.426	16.727	7.997	8.212	8.250	8.098	1.916	3.076	3.740	4.508
	NR	19.518	18.515	17.377	16.797	8.251	8.448	8.427	8.251	1.914	3.082	3.742	4.517
	CC	19.632	18.141	17.057	16.564	8.204	8.151	8.191	8.098	1.919	3.105	3.757	4.536
TRe	AA	15.217	17.004	16.616	16.594	7.639	8.226	8.160	8.189	1.940	3.070	3.773	4.520
	ID	15.392	17.151	16.717	16.605	7.644	8.266	8.192	8.195	1.939	3.062	3.765	4.511
	NR	13.226	15.496	15.724	15.522	6.797	7.528	7.738	7.649	1.955	3.092	3.794	4.550
	CC	15.186	16.943	16.585	16.595	7.638	8.208	8.148	8.188	1.941	3.074	3.775	4.521
TRf	AA	17.260	15.617	15.373	16.671	8.386	7.810	7.774	7.998	1.962	3.224	3.861	4.680
	ID	16.815	15.574	15.417	16.763	8.270	7.803	7.764	8.009	1.961	3.224	3.867	4.681
	NR	19.149	16.279	15.738	17.112	8.785	8.087	7.968	8.300	1.945	3.204	3.849	4.651
	CC	17.130	15.613	15.375	16.671	8.354	7.809	7.775	7.999	1.962	3.224	3.860	4.679

and MAE, and highest positive correlation coefficients. For stream flow data in Table 3, each station also gave different methods. The NR method is the most suitable method for TSc, TSd and TSe stations. Meanwhile the best method for TSb and TSf stations are coefficient of correlation (CC). AA method is the last choice method to estimate missing rainfall and stream flow data because it gave high values for RMSE and MAE, and low value for R test.

Fig. 2 shows the example comparison of RMSE, MAE and R method with various percentages of missing values for TRa and TRc stations. It proved that NR method gave the lowest value of RMSE and MAE errors for TRa station. The example plots of RMSE, MAE and R methods for stream flow data are shown in Fig. 3. It is clearly shown that CC is the best method for TSb station. Meanwhile for TRc station, NR method shows better results which gave lowest value of RMSE and MAE compared to other methods. Other than that, there are slightly increase in the value of correlation coefficient (R) for each estimation method for all stations.

However, after four interpolation methods could not be used because there is no information from neighboring stations, we decided to use the mean on the same day and month with different years as estimation of the missing values on that particular date. Table 4 is a snapshot of original data for rainfall (TRa) station and stream flow (TSa) station. From Table 4, the 66th row represent the missing item data for TRa station in day 66 (2005). Meanwhile, for TSa station, day 336 (1995) represent 336th row missing item value for stream flow data. We used mean of the same day and month with different years to replace these missing values. Table 5 presents the snapshot of few data using mean of the same day and month with different years. NA value of the rainfall data in day 66 (2005) is replaced by the mean of the non missing rainfall data in day 66 from 1995 until 2014.

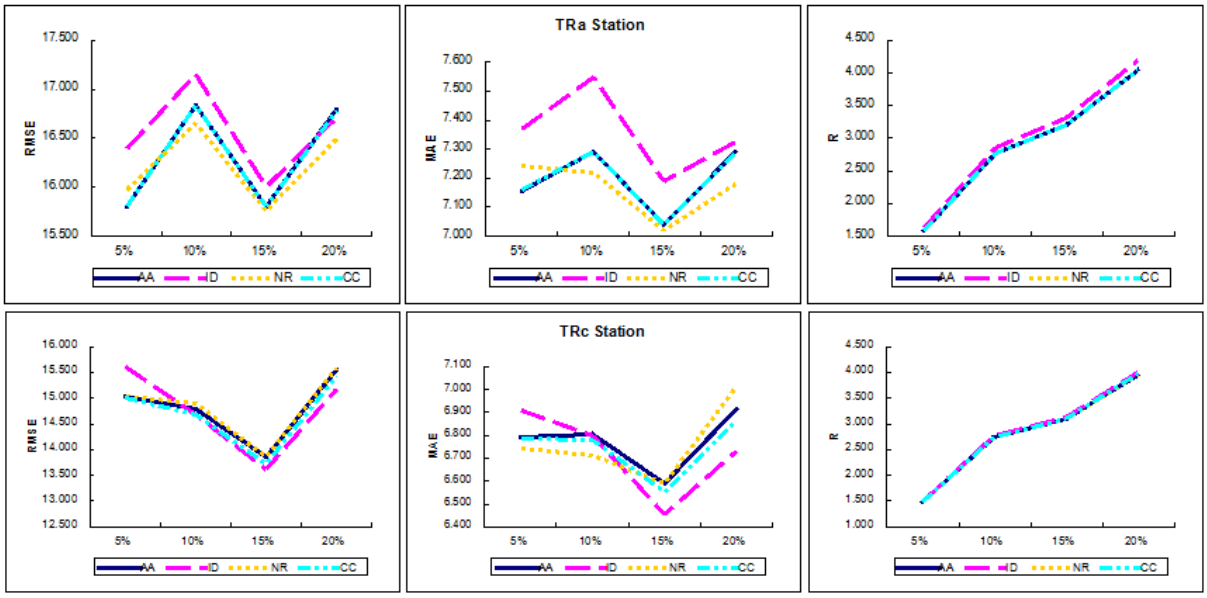

Fig. 2 Comparison of RMSE, MAE and R method with various percentages of missing values for TRa and TRc Stations.

Table 3 Comparison of estimation methods based on RMSE, MAE and R with four different percentages of missing values for stream flow data.

Station	Methods	RMSE				MAE				R			
		5%	10%	15%	20%	5%	10%	15%	20%	5%	10%	15%	20%
TSa	AA	32.592	33.817	34.030	33.910	16.687	17.079	17.485	17.750	2.879	4.098	4.831	5.494
	ID	26.509	28.045	28.869	28.705	12.820	13.299	13.831	13.995	2.713	3.826	4.512	5.117
	NR	31.014	31.423	32.917	32.934	17.865	18.090	18.476	18.401	2.764	3.848	4.751	5.409
	CC	30.026	31.750	32.565	32.327	15.085	15.730	16.491	16.659	2.806	3.993	4.740	5.375
TSb	AA	27.646	24.706	29.154	27.364	11.486	11.004	11.363	10.866	2.784	3.921	4.595	5.189
	ID	29.284	26.058	30.592	28.711	12.176	11.576	11.917	11.388	2.837	4.003	4.677	5.284
	NR	25.643	24.293	27.518	26.460	11.507	11.053	11.520	11.199	2.840	4.087	4.627	5.255
	CC	25.857	23.474	28.044	26.278	10.851	10.565	11.033	10.575	2.729	3.844	4.532	5.112
TSc	AA	46.762	44.115	46.733	45.191	24.506	23.958	24.548	24.251	2.762	3.761	4.467	5.045
	ID	46.110	42.327	46.208	44.173	22.879	22.111	22.660	22.184	2.832	3.948	4.406	4.976
	NR	7.093	7.048	7.159	6.924	3.906	3.816	3.797	3.684	2.993	4.130	4.813	5.449
	CC	46.669	43.973	46.699	45.130	24.388	23.852	24.500	24.181	2.768	3.774	4.466	5.045
TSd	AA	147.349	153.086	159.313	157.422	103.368	102.538	104.066	102.272	2.666	3.761	4.682	5.237
	ID	152.059	170.948	173.066	171.653	69.119	72.441	72.574	70.428	3.216	4.708	6.015	6.714
	NR	106.584	117.657	120.142	120.709	59.067	61.526	61.510	59.968	2.937	4.408	5.618	6.358
	CC	145.280	152.928	158.725	156.927	96.150	100.458	99.328	98.109	2.724	3.784	4.759	5.310
TSe	AA	202.912	226.334	223.323	222.337	168.531	173.758	172.091	169.680	2.352	3.667	4.635	5.301
	ID	195.296	219.717	216.743	215.791	155.170	160.582	158.865	156.477	2.418	3.746	4.740	5.414
	NR	25.494	20.400	21.739	19.861	8.750	7.497	8.107	7.473	2.863	4.361	5.547	6.306
	CC	195.926	221.608	217.129	216.728	156.360	164.571	159.710	158.521	2.412	3.723	4.734	5.398
TSf	AA	232.486	234.211	237.648	235.670	216.813	216.544	219.127	216.921	3.456	4.992	6.246	7.023
	ID	249.016	256.947	258.046	256.268	233.689	236.024	236.798	234.429	3.441	5.847	7.380	8.439
	NR	489.581	451.231	474.084	463.771	222.155	212.585	220.115	212.426	2.692	3.978	4.971	5.621
	CC	231.024	231.100	234.832	233.034	215.179	213.600	216.348	214.484	3.443	4.893	6.104	6.871

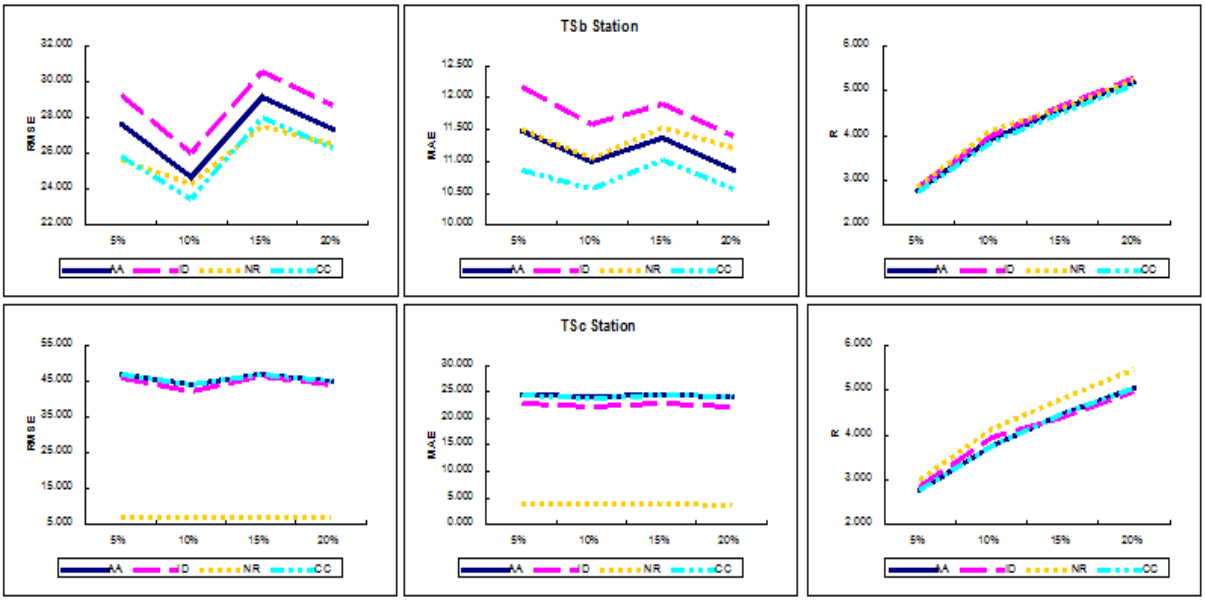

Fig. 3 Comparison of RMSE, MAE and R method with various percentages of missing values for TSb and TSc Stations.

Table 4 The snapshot of original data of rainfall (TRa) and stream flow (TSa).

Rainfall (TRa)		Stream flow (TSa)	
Day	Year (2005)	Day	Year (1995)
66	NA	336	NA
67	NA	337	NA
68	NA	338	NA
69	NA	339	NA
70	NA	340	NA

Table 5 The snapshot of rainfall (TRa) and stream flow (TSa) after using the mean of the same day and month with different years.

Rainfall (TRa)		Stream flow (TSa)	
Day	Year (2005)	Day	Year (1995)
66	8.8211	336	67.2313
67	6.3632	337	93.0913
68	9.2632	338	93.7327
69	4.6789	339	86.5873
70	14.9421	340	68.3453

CONCLUSION

In estimating missing rainfall and stream flow data, the Arithmetic Average (AA) method, Inverse Distance (ID) method, Normal Ratio (NR) method and Coefficient of Correlation (CC) method were compared. All of these methods have been tested at four different percentages of missing data (5%, 10%, 15%, 20%). The results which gave the minimum RMSE and MAE as well as highest positive correlation coefficients was chosen as the best method.

From six stations of the rainfall data, three stations recorded same technique which is ID method. Meanwhile, CC method is the second best method for all rainfall stations. For stream flow data, NR method are found to be the best estimation method among all especially TSc station which showed big difference values compared to the other three methods. It can be seen that most stations do not possess similar best method of choosing data treatment. However, the mean on the same day and month but at different years is taken to estimate the missing value on that particular date if there is no information from neighboring stations.

For future study, it is recommended to consider application of functional data analysis for the treatment of missing rainfall and stream flow data. Other suggestion for this study is to increase the number of neighboring stations involved as well as the distance between target station and nearby station can be increased until radius range of 10 km to 100 km in order to estimate better results for missing rainfall and stream flow data.

ACKNOWLEDGEMENT

This work was financially supported by the Ministry of Higher Education Malaysia (KPT MyBrain15).

REFERENCES

Burrough, P. A., McDonnell, R. A. 1998. *Principles of Geographical Information Systems*. Oxford: Oxford University Press.

Chow, V.T., Maidment, D. R. and Mays, L. W. 1988. *Applied Hydrology*. Singapore: Mc Graw Hill Book Company.

Dastorani, M. T, Moghadamnia, A., Piri, J. and Rico-Ramirez, M. A. 2010. Application of ANN and ANFIS models for reconstructing missing flow data. *Environmental Monitoring and Assessment*, 166, 421-434.

De Silva, R. P., Dayawansa, N. D. K., Ratnasari, M. D. 2007. A comparison of methods used in estimating missing rainfall data. *The Journal of Agricultural Science,* 3(2), 101-108.

Hasan, M. M., Croke, B. F. W. 2013. Filling gaps in daily rainfall data: a statistical approach. *20th International Congress on Modelling and Simulation*. 1–6 December 2013. Adelaide, Australia, 380-386.

Ibrahim, N., Wibowo, A. 2014. Support vector regression with missing data treatment based variables selection for water level prediction of Galas River in Kelantan Malaysia. *WSEAS Transactions on Mathematics*, 13, 69-78.

Paulhus, J. L. H., Kohler, M. A. 1952. Interpolation of Missing precipitation records. *Monthly Weather Review*, 80, 8, 129-133.

Suhaila, J., Deni, S. M., Jemain, A. A. 2008. Revised spatial weighting methods for estimation of missing rainfall data. *Asia-Pacific Journal of Atmospheric Sciences*, 44, 2, 93-104.

Teegavarapu, R. S. V., Chandramouli, V. 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. Journal of *Hydrology*, 312, 1, 191-206.