ASM Sc. J., 13, 2020
https://doi.org/10.32802/asmscj.2020.sm26(1.20)

Simulation Study of Adjusted Spatial Weighting Method to Estimate Missing Rainfall Data

Muhammad Az-zuhri Azman*, Roslinazairimah Zakaria and Siti Zanariah Satari

Faculty of Industrial Sciences & Technology, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300, Gambang, Kuantan, Pahang, Malaysia

Missing value especially in environmental study is a common problem including in rainfall modelling. Incomplete data will affect the accuracy and efficiency in any modelling process. In this study, simulation method is used to demonstrate the efficiency of the old normal ratio inverse distance correlation weighting method (ONRIDCWM) in solving missing rainfall data. The simulation study is used to identify the best parameters for correlation power of p, percentage of missing value and sample size, n of the ONRIDCWM through simulating for 10,000 times by varying the value of the parameters systematically. The results of the simulation are compared with other available weighting methods. The estimated complete rainfall data of the target station are compared and assessed with the observed data from the neighbouring station using mean, estimated bias (EB) and estimated root mean square error (ERMSE). The results show that ONRIDCWM is better than the other weighting methods for the correlation power of p at least four. For illustration of the weighting method, monthly rainfall data from Pahang is used to demonstrate the efficiency of the method using three error indices: S-Index, mean absolute error (MAE) and correlation, R.

I. INTRODUCTION

Missing values in rainfall data is a common and unavoidable problem faced during statistical analysis. The reason of having missing rainfall data are due to the human error when measuring the rainfall amount, malfunction of the instrument for a certain period of time especially during the extreme rainfall events, unsystematic way when storing the rainfall data or relocation of meteorological rainfall station. Thus, the problem of having missing value required an appropriate method or technique to handle it effectively as it reduces the statistical power of a study which can produce the biased estimates and can have a significant effect on the conclusions that can be drawn from the rainfall data(Ahmad Radi *et al.,* 2015; Azman *et al.,* 2015; Hasana & Crokea, 2013).

Basically, the procedures to handle missing rainfall values can be divided into three major classes as deterministic, stochastic and artificial intelligence based methods (Campozano *et. al.,* 2014). In this study the deterministic approach is applied. Three advantages of using deterministic approaches are robustness, easy for implementation and computationally efficient (Caldera *et. al.,* 2016; Campozano *et. al.,* 2014; Silva *et. al.,* 2007). The deterministic approach is based on mathematical models which considering certain factors such as distance and correlation for imputing the missing rainfall data. However, the best selection method for estimating missing rainfall values can be varied for different regions depending on their rainfall patterns, spatial and temporal distributions.

In this study, simulation method is used to determine the efficiency of old normal ratio inverse distance correlation weighting method (ONRIDCWM) proposed by Azman *et. al.,* (2015) in completing the missing rainfall data of a selected target station. All the results of ONRIDCWM are compared

*Corresponding author's e-mail: putra_autumn86@yahoo.com

to other weighting method by Suhalia *et. al.,* (2008) using the performance indicator of mean, estimated bias (EB) and estimated root mean square errors (ERMSE).

II. ESTIMATION WEIGHTING METHOD

The formulae of weighting methods suggested by Suhalia *et al.*, (2008)(CCWM, CCIDWM, NRIDWM and ONRIDWM) are as follows:

Correlation Coefficient Weighting Method (CCWM)

$$W_i = \frac{\rho_{it}^{p}}{\sum_{\substack{i=1 \\ i\neq 1}}^{N} \rho_{it}^{p}} \tag{1}$$

where ρ_{it}^{p} is the correlation coefficient between the target station S_t and the i^{th} neighbouring stations with $p = 2$.

Correlation Coefficient with Inverse Distance Weighting Method (CCIDWM)

$$W_i = \frac{\rho_{it}^{p} d_{it}^{-2}}{\sum_{\substack{i=1 \\ i\neq 1}}^{N} \rho_{it}^{p} d_{it}^{-2}} \tag{2}$$

where ρ_{it}^{p} is the coefficient of correlation between the target station t and the i^{th} neighbouring stations with $p = 2$ while d_{it} is the distance between the target station S_t and the i^{th} neighbouring station.

Normal Ratio with Inverse Distance Weighting Method (NRIDWM)

$$W_i = \frac{(n_i - 2)\rho_{it}^{2}(1-\rho_{it}^{2})^{-1} d_{it}^{-2}}{\sum_{\substack{i=1 \\ i\neq 1}}^{N} (n_i - 2)\rho_{it}^{2}(1-\rho_{it}^{2})^{-1} d_{it}^{-2}} \tag{3}$$

where ρ_{it}^{2} is the square of coefficient of correlation of daily rainfall data between the target station, S_t and i^{th} neighbouring station; n_i is the length of data or number of points that are used to compute the correlation coefficient; d_{it} is the distance between the target station, S_t and the i^{th} neighbouring station and W_i is the resultant weight.

Old Normal Ratio with Inverse Distance Weighting Method (ONRIDWM)

$$W_i = \frac{\frac{\mu_t}{\mu_i} d_{it}^{-2}}{\sum_{\substack{i=1 \\ i\neq 1}}^{N} \frac{\mu_t}{\mu_i} d_{it}^{-2}} \tag{4}$$

where d_{it} is the distance between the target station and the i^{th} neighbouring station while μ_t and μ_i are the sample mean of available data at the target station S_t and the i^{th} neighbouring station respectively and W_i is the resultant weight.

A. Modification of Estimation Weighting Method

1. Old Normal Ratio with Inverse Distance and Correlation Weighting Method (ONRIDCWM)

The weighting method proposed by Azman *et al.*, (2015) is known as old normal ratio with inverse distance and correlation weighting method (ONRIDCWM) and given by

$$W_i = \frac{\rho_{it}^{p} \frac{\mu_t}{\mu_i} d_{it}^{-2}}{\sum_{\substack{i=1 \\ i\neq 1}}^{N} \rho_{it}^{p} \frac{\mu_t}{\mu_i} d_{it}^{-2}} \tag{5}$$

where W_i is the respective weight, ρ_{it}^{p} is the correlation coefficient of power p, μ_t and μ_i are the sample mean of available data at the target station t and the i^{th} neighbouring station respectively and d_{it} is the distance between S_t and the i^{th} neighbouring station respectively.

III. SIMULATION STUDIES

In this simulation study, program is written using R language. Five different methods considered will be referred as CCWM, CIDWM, NRIDWM, ONRIDWM and ONRIDCWM. In the simulation process, the power of correlation, p for ONRIDCWM formula is varied from $p=1,2,3,4,5.$ For each sample, different percentages of missing values (PMV) are assigned: 5%, 15% and 30%, respectively. The estimated rainfall amount (X_t) is calculated using

$$X_t = \sum_{\substack{i=1 \\ i\neq 1}}^{N} W_i X_i \tag{6}$$

where X_t is the estimated value of the missing data at the target station S_t; N is the number of neighbouring stations; X_i is the observation at the i^{th} neighbouring station and W_i is the weight of the i^{th} neighbouring station with constraint $\sum_{i=1}^{N} W_i = 1.$

The data is fitted to the gamma model of two parameters, shape (α) and scale (β) and the pdf is given by

$$f(x;\alpha,\beta) = \frac{\beta^{-\alpha}x^{\alpha-1}}{\Gamma(\alpha)}\exp\left(\frac{-x}{\beta}\right), \tag{7}$$
$$\alpha>0, \beta>0, x>0.$$

The simulation process are repeated for 10,000 times, and the values of X are drawn from $X \sim \Gamma(0.7, 335)$ with different sample sizes, $n = 200, 600, 1200.$ The sets of synthetic data generated are compared and assessed using the performance indicators; mean of the estimated parameters ($\hat{\theta}$), estimated bias (EB) and estimated root mean square errors (ERMSE). The formula for mean of the estimated parameters are given by

$$\text{Mean of } \hat{\theta}: \bar{\hat{\theta}} = \frac{1}{simu}\sum \hat{\theta}_j \tag{8}$$

where *simu* is the number of simulations. So, in this case the value of *simu* is 10000. The parameters in the study are α and β. These values of mean are used to calculate the EB for both parameters of α and β. The formula of EB is given by

$$\text{EB of } \hat{\theta} = \left|\bar{\hat{\theta}} - \theta\right| \tag{9}$$

ERMSE is used to measure the difference between the predicted values by a model and the observed values. The ERMSE formula of the parameters is

$$\text{ERMSE of } \hat{\theta} = \sqrt{\frac{1}{simu}\sum\left(\hat{\theta}_j - \theta\right)^2} \tag{10}$$

The smallest value of EB and ERMSE will be chosen and this demonstrate that the estimated parameters are good.

IV. RESULTS AND DISCUSSION

This study compares the performance of weighting method proposed by Suhaila *et al.,* (2008) and Azman *et al.,* (2015), refer to Table 2 and Table 3. It is observed that, when the power of correlation coefficient is $p=4$ and $p=5$, as the number of sample sizes increasing, the mean of the estimated shape and scale parameters approach the true shape and scale parameters and the error assessment using EB and ERMSE become smaller. In contrast, when the percentage missing values increasing, the mean of the estimated shape and scale parameters diverge from the true shape and scale parameters and the error of EB and ERMSE is increasing. Thus, ONRIDCWM with correlation power of $p=4$ and $p=5$ outperform the other methods for all different percentages of missing values and sample sizes. It is also noted that, ONRIDWM is considered less superior than other methods by Suhaila *et al.*, (2008). However, by including the correlation of power, *p* the results of ONRIDCWM improved the efficiency when $p=4$ and $p=5$.

V. ILLUSTRATIVE EXAMPLES USING RAINFALL DATA FROM PAHANG STATIONS

In this section, rainfall data from Pahang was selected to show the performance of the proposed method ONRIDCWM by Azman *et. al.,* (2015) and compare with Suhaila *et al.,* (2008) methods. A good performance of weighting method indicates the high value for both values of S-index and correlation (*R*) but low for mean absolute error (MAE) value.

Table 1. Comparison of estimation method based on S-Index, MAE and *R* for various percentages of missing rainfall data in Pahang

Methods	5%	15%	30%
	S-Index		
CCWM	0.847	0.842	0.827
CIDWM	0.798	0.790	0.783
NRIDWM	0.817	0.809	0.794
ONRIDWM	0.808	0.803	0.786
ONRIDCWM	0.856	0.845	0.831
	MAE		
CCWM	85.519	85.820	86.057
CIDWM	85.969	85.989	86.474
NRIDWM	85.819	85.910	86.309
ONRIDWM	85.889	85.938	86.391
ONRIDCWM	85.447	85.798	85.882
	R		
CCWM	0.785	0.763	0.753
CIDWM	0.756	0.703	0.658
NRIDWM	0.769	0.730	0.688
ONRIDWM	0.763	0.724	0.673
ONRIDCWM	0.788	0.778	0.769

In this study, it is observed that from Table *1*, ONRIDCWM using the correlation power of $p = 4$ give the best results as compared to other methods from Suhaila *et al.*, (2008). However, when the number of sample sizes is increasing, the performance of each estimation method tends to decrease slightly in S-index and *R* but to increase slightly for MAE. Thus, this results shows that the proposed weighting method by Azman *et al.*, (2015) able to improve the existing weighting methods.

VI. CONCLUSION

In this study, simulation method is used to demonstrate the efficiency of the old normal ratio inverse distance correlation weighting method (ONRIDCWM) in solving missing rainfall data. The ONRIDCWM formula includes the correlation coefficient of power *p* whereas the old normal ratio inverse distance (ONRIDWM) formula suggested by Suhaila *et. al.,* (2008) did not include the correlation coefficient of power *p* term. Analysis using simulation study found that the ONRIDCWM formula is able to provide optimal result when the power of correlation is four and five with minimum errors. Hence, ONRIDCWM improved the results of ONRIDWM by Suhaila *et. al.,* (2008). The results also show that ONRIDCWM is better than the other weighting methods for the correlation power of *p* at least four. For future study, the viability of the suggested method ONRIDCWM can be tested using other variations of factors including missing values more than 30%.

VII. ACKNOWLEDGEMENTS

The authors would like to convey thanks to Universiti Malaysia Pahang for the financial support (RDU1703208). We thank the anonymous referees for their useful suggestions.

VIII. REFERENCES

Ahmad Radi, N. F., Zakaria, R., & Azman, M. A. 2015, Estimation of Missing Rainfall Data Using Spatial Interpolation and Imputation Methods. *AIP Conference Proceedings*, *1643*, 42–48.

Azman, M. A., Zakaria, R., & Ahmad Radi, N. F. 2015, Estimation of Missing Rainfall Data in Pahang Using Modified Spatial Interpolation Weighting Methods. *AIP Conference Proceedings*, *1643*, 65-72.

Caldera, H. P. G. M., Piyathisse, V. R. P. C., & Nandalal, K. D. W. 2016, A Comparison of Methods of Estimating Missing Daily Rainfall Data. *Engineer: Journal of the Institution of Engineers*, *49*(4), 1–8.

Campozano, L., Sánchez, E., Aviles, A., & Samaniego, E. 2014, Evaluation of Infilling Methods for Time Series of Daily Precipitation and Temperature: The case of the Ecuadorian Andes. *Maskana*, *5*(1), 99–115.

Hasana, M., & Crokea, B. 2013, Filling Gaps in Daily Rainfall Data: A Statistical Approach. *20th International Congress on Modelling and Simulation, Adelaide, Australia*, (December), 1–6.

Silva, R. P. De, Dayawansa, N. D. K., & Ratnasiri, M. D. 2007, A Comparison of Methods Used in Estimating Missing Rainfall Data, (May), 101–108.

Suhalia, J., Sayang, M. D., & Jemain, A. A. 2008, Revised Spatial Weighting Methods for Estimation of Missing Rainfall Data. *Asia-Pacific Journal of Atmospheric Sciences*, *44*(2), 93–104.

Table 2. Simulation results of performance indicator for mean of estimated shape parameter, $\bar{\hat{\alpha}}$ and error assessment (EB and ERMSE) with different power of correlations, number of sample sizes and percentages of missing value ($\alpha = 0.7$)

Performance Indicator	No. of Sample Size		200			600			1200		
	Percentage of missing values (%)		5	15	30	5	15	30	5	15	30
Mean of $\hat{\alpha}$	CCWM		0.713	0.723	0.742	0.707	0.718	0.737	0.706	0.716	0.735
	CIDWM		0.719	0.743	0.785	0.713	0.739	0.779	0.712	0.737	0.777
	NRIDWM		0.717	0.736	0.769	0.711	0.731	0.764	0.710	0.729	0.764
	ONRIDWM		0.718	0.740	0.779	0.712	0.736	0.771	0.712	0.734	0.771
	p - power of ONRIDCWM	1	0.717	0.736	0.770	0.711	0.731	0.764	0.710	0.729	0.764
		2	0.715	0.729	0.756	0.709	0.724	0.751	0.708	0.723	0.750
		3	0.714	0.725	0.746	0.707	0.719	0.740	0.707	0.718	0.739
		4	**0.712**	**0.721**	**0.737**	**0.706**	**0.715**	**0.732**	**0.705**	**0.714**	**0.730**
		5	**0.711**	**0.717**	**0.731**	**0.705**	**0.712**	**0.726**	**0.704**	**0.711**	**0.724**
EB	CCWM		0.013	0.023	0.042	0.007	0.018	0.037	0.006	0.016	0.035
	CIDWM		0.019	0.043	0.085	0.013	0.039	0.079	0.012	0.037	0.077
	NRIDWM		0.017	0.036	0.069	0.011	0.031	0.064	0.010	0.029	0.064
	ONRIDWM		0.018	0.040	0.079	0.012	0.036	0.071	0.012	0.034	0.071
	p - power of ONRIDCWM	1	0.017	0.036	0.070	0.011	0.031	0.064	0.010	0.029	0.064
		2	0.015	0.029	0.056	0.009	0.024	0.051	0.008	0.023	0.050
		3	0.014	0.025	0.046	0.007	0.019	0.040	0.007	0.018	0.039
		4	**0.012**	**0.021**	**0.037**	**0.006**	**0.015**	**0.032**	**0.005**	**0.014**	**0.030**
		5	**0.011**	**0.017**	**0.031**	**0.005**	**0.012**	**0.026**	**0.004**	**0.011**	**0.024**
ERMSE	CCWM		0.064	0.069	0.088	0.037	0.044	0.068	0.027	0.037	0.063
	CIDWM		0.066	0.078	0.113	0.039	0.055	0.093	0.029	0.048	0.088
	NRIDWM		0.065	0.076	0.104	0.038	0.052	0.084	0.028	0.043	0.080
	ONRIDWM		0.065	0.078	0.111	0.038	0.054	0.090	0.028	0.046	0.085
	p - power of ONRIDCWM	1	0.065	0.076	0.104	0.038	0.052	0.084	0.028	0.043	0.080
		2	0.064	0.073	0.096	0.037	0.048	0.076	0.027	0.040	0.071
		3	0.064	0.070	0.090	0.037	0.046	0.069	0.027	0.037	0.065
		4	**0.064**	**0.069**	**0.086**	**0.037**	**0.044**	**0.065**	**0.026**	**0.035**	**0.059**
		5	**0.063**	**0.067**	**0.083**	**0.036**	**0.043**	**0.061**	**0.026**	**0.034**	**0.055**

Table 3. Simulation results of performance indicator for mean of estimated scale parameter, $\overline{\hat{\beta}}$ and error assessment (EB and ERMSE) with different power of correlations, number of sample sizes and percentages of missing value ($\beta = 355.0$)

Performance Indicator	No. of Sample Size		200			600			1200		
	Percentage of missing values (%)		5	15	30	5	15	30	5	15	30
Mean of $\hat{\beta}$	CCWM		330.295	323.717	311.312	330.856	323.598	312.876	329.910	325.241	312.800
	CIDWM		327.518	315.305	294.266	327.814	314.517	295.828	327.114	315.648	295.921
	NRIDWM		328.390	318.168	300.838	328.936	317.780	301.743	328.091	318.617	301.753
	ONRIDWM		327.752	316.335	296.867	328.341	315.990	298.237	327.586	316.428	298.829
	p - power of ONRIDCWM	1	328.350	318.042	300.610	328.919	317.739	301.688	328.079	318.590	301.714
		2	329.213	320.616	305.690	329.851	320.535	307.241	328.970	321.536	307.036
		3	329.913	322.941	310.012	330.654	322.864	311.800	329.694	323.965	311.563
		4	**330.469**	**324.777**	**313.420**	**331.275**	**324.618**	**315.282**	**330.257**	**325.827**	**315.062**
		5	**330.903**	**326.210**	**316.035**	**331.747**	**325.965**	**317.913**	**330.700**	**327.250**	**317.789**
EB	CCWM		4.705	11.283	23.688	4.144	11.402	22.124	5.090	9.759	22.200
	CIDWM		7.482	19.695	40.734	7.186	20.483	39.172	7.886	19.352	39.079
	NRIDWM		6.610	16.832	34.162	6.064	17.220	33.257	6.909	16.383	33.247
	ONRIDWM		7.248	18.665	38.133	6.659	19.010	36.763	7.414	18.572	36.171
	p - power of ONRIDCWM	1	6.650	16.958	34.390	6.081	17.261	33.312	6.921	16.410	33.286
		2	5.787	14.384	29.310	5.149	14.465	27.759	6.030	13.464	27.964
		3	5.087	12.059	24.988	4.346	12.136	23.200	5.306	11.035	23.437
		4	**4.531**	**10.223**	**21.580**	**3.725**	**10.382**	**19.718**	**4.743**	**9.173**	**19.938**
		5	**4.097**	**8.790**	**18.965**	**3.253**	**9.035**	**17.087**	**4.300**	**7.750**	**17.211**
ERMSE	CCWM		40.668	41.723	49.551	23.681	27.130	35.898	17.276	20.174	32.272
	CIDWM		40.850	43.219	56.456	24.077	30.713	45.956	18.053	25.232	43.186
	NRIDWM		40.736	42.847	53.078	23.782	29.697	42.503	17.751	23.702	39.463
	ONRIDWM		40.760	43.122	54.805	23.907	30.531	45.099	17.928	24.869	41.688
	p - power of ONRIDCWM	1	40.748	42.847	53.211	23.779	29.717	42.527	17.753	23.717	39.484
		2	40.709	42.325	51.057	23.646	28.425	39.163	17.500	22.173	35.806
		3	40.712	41.964	49.441	23.604	27.434	36.524	17.313	20.959	32.882
		4	**40.717**	**41.684**	**48.360**	**23.584**	**26.787**	**34.474**	**17.160**	**20.111**	**30.630**
		5	**40.711**	**41.477**	**47.581**	**23.589**	**26.304**	**32.948**	**17.045**	**19.478**	**28.875**