

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

186,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



## Restoration of hydrological data in the presence of missing data via Kohonen Self Organizing Maps

Marlinda A. Malek\*, Siti Mariyam Shamsuddin\*\* and Sobri Harun\*\*  
*Universiti Tenaga Nasional\*, Universiti Teknologi Malaysia\*\*  
 Malaysia*

### 1. Introduction

The Malaysia National Network system utilises three methods of rainfall data collection, namely manual, chart recording and data logger method. These methods are simultaneously used at most rainfall stations. This leads to, where occurrence of missing data exists, the possibilities of missing data taken the shape of three predictable patterns. The missing data patterns identified are either in the form of missing data from one recording method or two recording methods or all three recording methods. It is also noted that, where data is available, there are prevalent measurement inconsistencies between the three methods, even though all apparatus are placed at the same rainfall station. Through data exploration exercise, it is found that the discrepancy between one method of measurement and another may range between 0% - 100%, indicating a relatively unstable data to be relied upon. The current practice to resolve the problem of missing data from one recording method is to substitute the missing data with the remaining recorded data. Similarly, if data from any two of the three recording methods are missing then the available data from the third method is used as a reference. In statistical terminology, this method of substitution is referred as "Hot-deck Imputation". While easily applied, the obvious drawbacks of this method, is the fact that it is not supported by any scientific rationale and it cannot be applied when data is not available from all three recording methods.

This study explores and analyzes techniques in patching daily rainfall records of selected rainfall stations in the states of Perlis, Selangor and Johor, as these states generally represent a fairly complete geographical coverage of Peninsular Malaysia i.e. northern, middle and southern, using single-value approach. The rationale for deployment of single-value approach is due to the localised characteristics of Malaysia's rainfall (Sani, 1986), and the availability of rainfall records from multiple methods of data collection at each rainfall station. Single-value approach is adopted to avoid inconsistent spatial correlation between two distance rainfall stations. A study by Tang *et. al.* 1996, showed that spatial correlation between two separate rainfall stations in Malaysia fluctuates inconsistently every month.

## 2. Model Conceptualisation

The proposed model is a hybrid of Unsupervised Artificial Neural Network and Nearest Neighbour Imputation techniques. The objective of the task is to “generalise” the said result in such a way that it is universally applicable to any given data set.

### 2.1 Unsupervised Artificial Neural Network (ANN)

In Unsupervised ANN learning, the objective is to find the natural structure inherent in the input data. There are a number of unsupervised learning schemes, including competitive learning, adaptive resonance theory and Self-Organising feature Maps (SOMs). A well known type of SOM is Kohonen Network, implemented in this study. Kohonen Network maps input vectors (patterns) of arbitrary dimension  $N$  onto a discrete map with 1 or 2 dimensions. Patterns close to one another in the input space should be close to one another in the map, and they should be topologically ordered. A Kohonen Network is comprised of a grid of output units and  $N$  input units. The input pattern is fed to each output unit. The input lines to each output unit are weighted. These weights are initialised to small random numbers. The winning output unit is simply the unit with the weight vector that has the smallest Euclidean distance to the input pattern. The neighbourhood of a unit is defined as all units within some distance of that unit on the map (not in weight space). The weights of every unit in the neighbourhood of the winning unit (including the winning unit itself) are updated. This will move each unit in the neighbourhood closer to the input pattern. As time progresses the learning rate and the neighbourhood size are reduced. If the parameters are well chosen, the final network should capture the natural clusters in the input data.

### 2.2 Nearest Neighbour (NNeigh) Imputation technique

The NNeigh model has the advantage of simple computational approach. It represents only the most similar situations corresponding to the current situation. The basic concept of this model lies in the fact that similar situations will lead to similar outcomes. Thus the nearest-neighbour technique looks into the history of the events in the past data. The similarity of the present situation with the past ones is defined in terms of similarity metric. In addition, this technique achieves consistently high performance without *a priori* assumptions about the data distributions from which the training samples are drawn. (Porporato and Ridolfi, 1997; Sivakumar et al. 1999). The predictive model adopted here is based on the fact that Malaysia's rainfall is localised, where rainfall variability is more significant in spatial compared to temporal. The localised weather feature provides the notion that some days are similar. This study, capitalizes on this notion to select a probable day to impute missing data using distance measure criteria.

## 3. Data Preparation

The accuracy of simulated data using SOM technique is affected by sample size and percentage of data not present in the training set. To ensure that the proposed method is robust enough to cope with the vagaries due to sample size and extreme data insufficiency, five hypothetical sets of incomplete data are created with 20%, 40%, 60%, 70% and 80% of the days with missing data.

3.1 SOM Normalization

In determining the SOM structure, during normalization, analyses were done using two normalization methods, namely Linear Scaling to Unit Variance (Linear Normalization) and Rao Method (Rao Normalization). Below are the results obtained using Linear Normalization compared to Rao Normalization at rainfall station 3117070 in Selangor. Here, hypothetical data set of 80% days with missing data was used. The proportion of missing entries is varied to 40% days with up to two missing entries per day, while the other 40% of the days comprised of completely missing entries for all three recording methods. Table 1 exhibits part of the results on rainfall station 3117070 from manual data collection. Data used was from 1.2.2000 till 28.9.2002. The first column in Table 1 is the original data from the manual method of data recording. The imputed data obtained using Rao Normalization and Linear Normalization is as shown in the second and third column respectively.

Manual-original data	Manual-SOM- Rao Normalization	Manual-SOM- Linear Normalization
8	8.03501	8.173701672
9	8.97422	8.332597855
0	4.40E-11	0.000139165
1	0.962104	0.844107669
0	4.40E-11	0.000139165
0	4.40E-11	0.000139165
0	4.40E-11	0.000139165
0	4.40E-11	0.000139165
11	10.9827	10.94010085
0	4.40E-11	0.000616132
0	4.40E-11	0.000616132
0	4.40E-11	0.000616132
0	4.40E-11	0.000139165
0	4.40E-11	0.000139165
0	4.40E-11	0.000139165
0	4.40E-11	0.000139165
0	4.40E-11	0.000139165
0	4.40E-11	0.000139165
70.5	70.4958	71.865947
0	4.40E-11	0.000139165
11.5	11.4993	11.46437728
0	4.40E-11	0.000139165
0	4.40E-11	0.000139165
21.5	21.3622	21.40545904
26.5	26.386	27.93244575

16.5	16.474	16.26039792
0	4.40E-11	0.000616132
0	4.40E-11	0.000616132
0	4.40E-11	0.000616132
0	4.40E-11	0.000616132
0	4.40E-11	0.000616132
0	4.40E-11	0.000616132
46	46.0474	47.47499025
19.5	19.456	18.81669778
31.5	31.4156	30.26038128
31	31.0374	30.26038128
0	4.40E-11	0.000616132
0	4.40E-11	0.000616132
71.5	71.5059	71.865947
0	4.40E-11	0.000616132
1.5	1.4767	1.49589821

Table 1. Comparison between Rao and Linear Normalization in SOM structure at rainfall station 3117070 with 80% of days with missing data (Manual method of recording).

It is found that imputed values using Rao Normalization are closer to the original values as compared to Linear Normalization method.

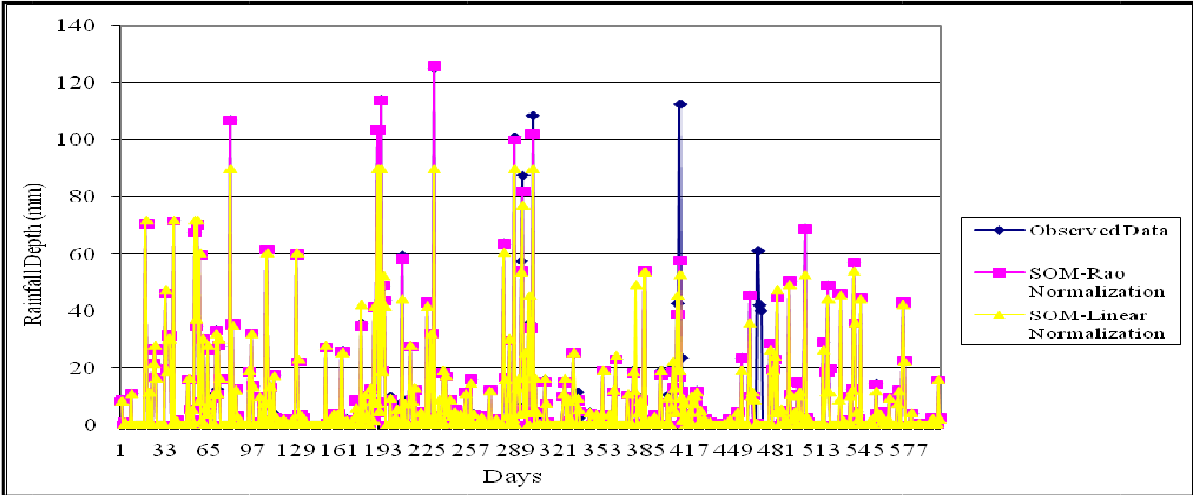


Fig. 1. Comparison between original or observed data and the imputed values obtained from Rao Normalization and Linear Normalization methods in SOM structure at rainfall station 3117070 (Manual method of recording)

Figure 1 shows that Rao Normalization peak values mostly coincides whereas Linear Normalization peak values mostly do not coincides with the peaks of the original data. Thus, in this study, the Rao Normalization method is preferred over Linear Normalization method in SOM.

## 4. Model Evaluation

The model is typically evaluated by assessing the difference between the observed values and the predicted values computed. In this scenario, the proposed model is used to predict missing or unobserved values. Once all missing values are completely predicted then the evaluation takes place. Results using SOM Model before and after hybrid with NNNeigh technique are analysed separately. The strength of the models is evaluated by assessing:

- 4.1 Efficiency, (E) of SOM Model
- 4.2 Quantization Error and Topological Error on SOM Model
- 4.3 Graphical Presentation on SOM Model
- 4.4 Visualization Analysis on SOM Model

### 4.1 Efficiency of SOM Model

The accuracy of the imputed values obtained from the proposed model, is assessed based on the percentage of accuracy to the original data, using Coefficient of Efficiency, (E), (Aitken, 1974):

$$E = \frac{\sum (QA - \overline{QA})^2 - \sum (QA - Q)^2}{\sum (QA - \overline{QA})^2} \quad (1)$$

where,  $QA$  is the original value,

$Q$  is the imputed value,

$\overline{QA}$  is the mean of original values.

Value of 1.0 obtained from E exhibits a perfect efficiency of the proposed model.

### 4.2 Quantization Error and Topological Error on SOM Model

The quality of maps produced is measured, based on two indicators, namely the quantization error and topological error. The quantization error is given by:

$$\varepsilon_q = \frac{1}{N} \sum_{i=1}^N \|X_i - m_c\| \quad (2)$$

where,

$X_i$  input data vector

$N$  total number of input samples

$m_c$  best matching reference vector

The topological error is given by:

$$\varepsilon_t = \frac{1}{N} \sum_{k=1}^N u(X_k) \quad (3)$$

where,

$X_k$  input data vector

$N$  total number of input samples

$\mu$  matching unit

Table 2 to 8 presents the efficiency, quantization and topological errors obtained from the proposed models when dealing with variety percentages of missing data.

Percentage of days with hypothetical missing data	Rainfall Station	Methods of data collection	SOM Model Efficiency	Quantization Error	Topological Error	SOM-NNeigh Hybrid Model Efficiency
20% i.e. 10% missing up to two methods - 10% missing from all methods	6401002	Manual	0.9996	0.0310	0.0336	0.4314
		Chart Recorder	0.9996	0.0240		0.6177
		Data Logger	0.9995	0.0148		0.5864
	6402008	Manual	0.9999	0.0201	0.0339	0.9453
		Chart Recorder	0.9998	0.0124		0.9578
		Data Logger	0.9997	0.0149		0.8759
40% i.e. 30% missing up to two methods - 10% missing from all methods	6401002	Manual	0.9997	0.0599	0.0336	0.2338
		Chart Recorder	0.9989	0.0431		0.5342
		Data Logger	0.9993	0.0325		0.4392
	6402008	Manual	0.9998	0.0469	0.0339	0.8783
		Chart Recorder	0.9999	0.0598		0.7094
		Data Logger	0.9998	0.0499		0.1614

Table 2. Various percentage of missing data for the state of Perlis

Percentage of days with hypothetical missing data	Rainfall Station	Methods of data collection	SOM Model Efficiency	Quantization Error	Topological Error	SOM-NNeigh Hybrid Model Efficiency
60% i.e. 50% missing up to two methods - 10% missing from all methods	6401002	Manual	0.9992	0.0935	0.0336	0.1155
		Chart Recorder	0.9997	0.0611		0.5468
		Data Logger	0.9996	0.0612		0.3860
	6402008	Manual	0.9998	0.0669	0.0339	0.8728
		Chart Recorder	0.9999	0.0647		0.7071
		Data Logger	0.9996	0.0593		0.1433

70% i.e. 60% missing up to two methods - 10% missing from all methods	6401002	Manual	0.9995	0.1165	0.0336	0.0661
		Chart Recorder	0.9995	0.0851		0.5014
		Data Logger	0.9993	0.0846		0.3888
	6402008	Manual	0.9999	0.0782	0.0339	0.8722
		Chart Recorder	0.9998	0.0740		0.5470
		Data Logger	0.9995	0.0740		0.1292

Table 3. Various percentage of missing data for the state of Perlis

Percentage of days with hypothetical missing data	Rainfall Station	Methods of data collection	SOM - Model Efficiency	Quantization Error	Topological Error l	SOM-NNeigh Hybrid Model Efficiency
20% i.e. 10% missing up to two methods - 10% missing from all methods	3117070	Manual	0.9999	0.0131	0.0334	0.8701
		Chart Recorder	0.9999	0.0129		0.8740
		Data Logger	0.9998	0.0183		0.7534
	3411017	Manual	0.9997	0.0099	0.0335	0.8928
		Chart Recorder	0.9992	0.0152		0.8906
		Data Logger	0.9997	0.0184		0.8529
40% i.e. 30% missing up to two methods - 10% missing from all methods	3117070	Manual	0.9999	0.0251	0.0334	0.8705
		Chart Recorder	0.9999	0.0196		0.8740
		Data Logger	0.9999	0.0249		0.7536
	3411017	Manual	0.9998	0.0168	0.0335	0.8928
		Chart Recorder	0.9991	0.0435		0.8211
		Data Logger	0.9994	0.0373		0.8239

Table 4. Various percentage of missing data for the state of Selangor



Percentage of days with hypothetical missing data	Rainfall Station	Methods of data collection	SOM Model Efficiency	Quantization Error	Topological Error	SOM-NNeigh Hybrid Model Efficiency
60% i.e. 50% missing up to two methods - 10% missing from all methods	3117070	Manual	0.9999	0.0415	0.0334	0.8680
		Chart Recorder	0.9993	0.0337		0.8721
		Data Logger	0.9997	0.0430		0.6874
	3411017	Manual	0.9998	0.0319	0.0335	0.8880
		Chart Recorder	0.9995	0.0595		0.8015
		Data Logger	0.9996	0.0504		0.8144
70% i.e. 60% missing up to two methods - 10% missing from all methods	3117070	Manual	0.9997	0.0501	0.0334	0.8656
		Chart Recorder	0.9998	0.0413		0.8681
		Data Logger	0.9991	0.0547		0.6553
	3411017	Manual	0.9998	0.0386	0.0335	0.8656
		Chart Recorder	0.9995	0.0666		0.7597
		Data Logger	0.9996	0.0588		0.8027

Table 5. Various percentage of missing data for the state of Selangor

Percentage of days with hypothetical missing data	Rainfall Station	Methods of data collection	SOM Model Efficiency	Quantization Error	Topological Error	SOM-NNeigh Hybrid - Model Efficiency
20% i.e. 10% missing up to two methods - 10% missing from all methods	2232001	Manual	0.9996	0.0217	0.0335	0.6958
		Chart Recorder	0.9997	0.0206		0.7714
		Data Logger	0.9987	0.0157		0.6610
	2235163	Manual	0.9993	0.0218	0.0335	0.8425
		Chart Recorder	0.9992	0.0210		0.9090
		Data Logger	0.9991	0.0202		0.8750

40% i.e. 30% missing up to two methods – 10% missing from all methods	2232001	Manual	0.9993	0.0568	0.0335	0.6570
		Chart Recorder	0.9994	0.0662		0.7236
		Data Logger	0.9988	0.0496		0.2531
	2235163	Manual	0.9990	0.0646	0.0335	0.6590
		Chart Recorder	0.9991	0.0512		0.8938
		Data Logger	0.9997	0.0496		0.8416

Table 6. Various percentage of missing data for the state of Johor

Percentage of days with hypothetical missing data	Rainfall Station	Methods of data collection	SOM Model Efficiency	Quantization Error	Topological Error	SOM-NNeigh Hybrid Model Efficiency
60% i.e. 50% missing up to two methods - 10% missing from all methods	2232001	Manual	0.9993	0.0869	0.0335	0.4785
		Chart Recorder	0.9994	0.0983		0.7086
		Data Logger	0.9992	0.0942		-0.6085
	2235163	Manual	0.9993	0.1084	0.0335	0.5278
		Chart Recorder	0.9996	0.0801		0.8693
		Data Logger	0.9995	0.0811		0.8193
70% i.e. 60% missing up to two methods - 10% missing from all methods	2232001	Manual	0.9996	0.0997	0.0335	0.4801
		Chart Recorder	0.9994	0.1338		0.6428
		Data Logger	0.9991	0.1052		-0.7149
	2235163	Manual	0.9991	0.1431	0.0335	0.4088
		Chart Recorder	0.9996	0.0923		0.8587
		Data Logger	0.9994	0.0978		0.8083

Table 7. Various percentage of missing data for the state of Johor

Below are the general characteristics found at all rainfall stations in the three states analyzed based on Table 2 to 7.

- a) Values of topographic error are found to be similar at the same rainfall station despite different percentage of missing data applied.
- b) Values of topographic and quantization error produced are small.
- c) Magnitude of model efficiency (E) for SOM Model is found to be high as percentage of missing data increases. This shows the superiority of SOM Model.
- d) Nevertheless, efficiency of SOM Model after hybrid with NNeigh technique (SOM-NNeigh Hybrid) is the same or dropped lower.
- e) Efficiency value for SOM Model is at fixed value of 99% at all rainfall stations in the three states analysed, for all data correlation values. As such, it can be concluded that the clustering concept embedded in SOM are not affected by the quality of input data used.

In order to investigate the outcome beyond 70% of days with missing data, three hypothetical data sets comprising of 80% days with missing data are created. In the first hypothetical data set, 70% are days with up to two missing entries per day, while the other 10% of the days comprised of completely missing entries for all three recording methods. In the second hypothetical data set, the proportion is varied to 40% and 40% respectively and in the third hypothetical data set, the proportion is varied to 10% and 70% respectively.

80% of days with hypothetical missing data	Rainfall Station	Methods of data collection	SOM Model Efficiency	Quantization Error	Topological Error	SOM - Neigh Hybrid Model Efficiency
(70% missing up to two methods+ 10% missing from all three methods)	6401002	Manual	0.9998	0.1172	0.0335	0.6144
		Chart Recorder	0.9997	0.1190		0.2583
		Data Logger	0.9997	0.1059		0.5300
(40% missing up to two methods 40% missing from all three methods)	6401002	Manual	0.9995	0.0887	0.0504	0.1063
		Chart Recorder	0.9998	0.0771		0.0218
		Data Logger	0.9999	0.0672		0.0549
(10% missing up to two methods+ 70% missing from all three methods)	6401002	Manual	1.000	0.0465	0.1019	0.1411
		Chart Recorder	1.000	0.0340		0.0012
		Data Logger	1.000	0.0190		-0.0264

Table 8. Various percentage of missing data for the state of Perlis

80% of days with hypothetical missing data	Rainfall Station	Methods of data collection	SOM Model Efficiency	Quantization Error	Topological Error	SOM-NNeigh Hybrid Model Efficiency
(70% missing up to two methods + 10% missing from all three methods)	3411017	Manual	0.9999	0.0401	0.0335	0.8449
		Chart Recorder	0.9998	0.0685		0.8106
		Data Logger	0.9991	0.0588		0.8987
(40% missing up to two methods + 40% missing from all three methods)	3411017	Manual	0.9999	0.0479	0.0502	0.3080
		Chart Recorder	0.9995	0.0478		0.5124
		Data Logger	0.9996	0.0506		0.4070
(10% missing up to two methods + 70% missing from all three methods)	3411017	Manual	0.9999	0.0126	0.1008	-0.1269
		Chart Recorder	0.9999	0.0170		0.0612
		Data Logger	0.9998	0.0288		-0.0175

Table 9. Various percentage of missing data for the state of Selangor

80% of days with hypothetical missing data	Rainfall Station	Methods of data collection	SOM Model Efficiency	Quantization Error for SOM Model	Topological Error for SOM Model	SOM-NNeigh Hybrid Model Efficiency
(70% missing up to two methods + 10% missing from all three methods)	2235163	Manual	0.9994	0.0977	0.0335	0.6661
		Chart Recorder	0.9996	0.1087		0.4210
		Data Logger	0.9996	0.0972		0.3761
(40% missing up to two methods + 40% missing from all three methods)	2235163	Manual	0.9996	0.1203	0.0503	0.3733
		Chart Recorder	0.9992	0.0936		0.3527
		Data Logger	0.9994	0.0796		0.5339
(10% missing up to two methods + 70% missing from all three methods)	2235163	Manual	0.9972	0.0853	0.1012	-0.6269
		Chart Recorder	0.9998	0.0477		-0.7051
		Data Logger	0.9997	0.0712		-0.7905

Table 10. Various percentage of missing data for the state of Johor

Based on Table 8 to 10, below are findings on model efficiency (E), quantization and topological error found at all rainfall stations in Perlis, Selangor and Johor.

a) SOM Model maintains high values of E i.e. approximately 1.0, at all level of data correlations for all rainfall stations in the three states analysed. This again, exhibits the superiority of SOM method.

b) Based on E before SOM Model is hybrid with NNeigh, the proposed model proves to be reliable even when it is applied to records with large amount of missing data entries i.e. 80%. The mixtures of the missing data are deliberately varied to test the robustness of the proposed model. In all cases, E at 10% missing up to two methods and 70% missing up to two methods , shows values of approximately 1.0. This shows that efficiency of this model is high even when the amount of data missing up to two methods of recordings were increased from 10% to 70%.

c) After SOM is hybrid with NNeigh, the E value for all cases, exhibits values closer to 1.0, when 10% missing from all three methods compared to when 40% missing from all three methods. It seems that the limit of which this model could cope with amount of missing from all three methods is approximately 40% where values of E approaches to 0 or negative in most of the cases.

d) In this study, it is found that the values of final quantisation and topological error are close to 0 at all rainfall stations. This shows that, in all cases, the values of Best Matching Unit (BMU) obtained, approximately resemble the input data. If the configuration of the model has not yet reached the stable state in the learning process, both the quantization and topological error remains significantly high i.e. approximately 1.0.

4.3 Graphical Presentation on SOM Model

Since the true values of the ‘imaginary’ missing data are known, judgment on the goodness of the model can be explicitly made. Graphical presentation on the imputed hypothetical missing data compared to the original data at rainfall station 6401002 , Perlis for 20% to 80% of days with missing entries are presented in Figure 2a to 8b. The outcomes are distinguished based on colours and symbols as listed in Table 11.








Outcomes from SOM Model	predicted/imputed data	red	
	observed/original data	green	
Outcomes from SOM-NNeigh Hybrid model	predicted/imputed data	red	
	observed/original data	light blue	
Symbols		Data from different types of recording	
		Manual	
		Chart Recorder	
		Data Logger	

Table 11. Color coding and symbols for graphical presentations of various models for Figure 2a to 8b.

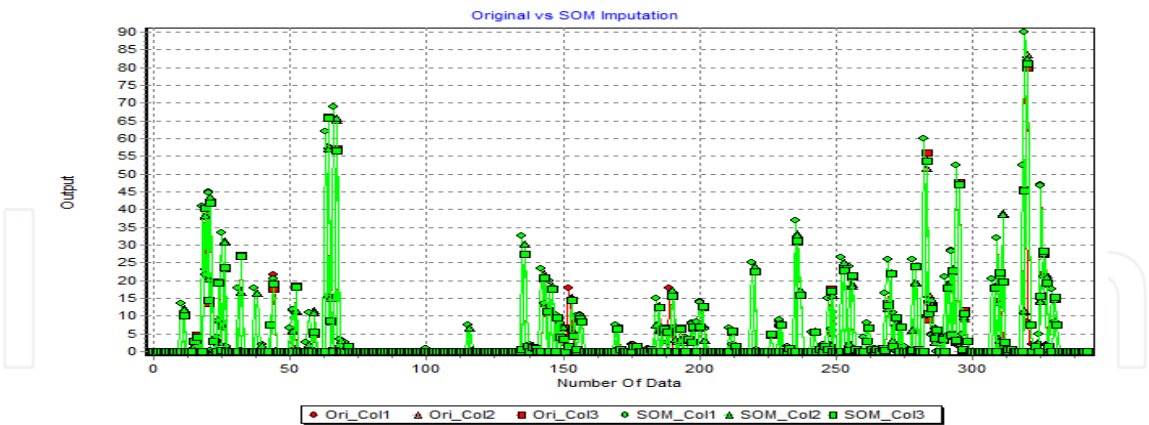


Fig. 2a. Comparison between original and imputed data using SOM Model for 20% days of hypothetical missing data set

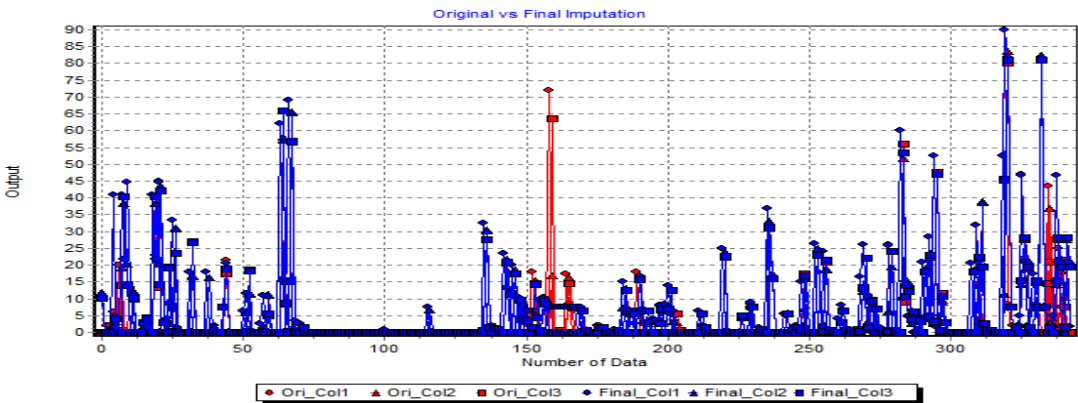


Fig. 2b. Comparison between original and imputed data using SOM-NNNeigh Hybrid Model for 20% days of hypothetical missing data set

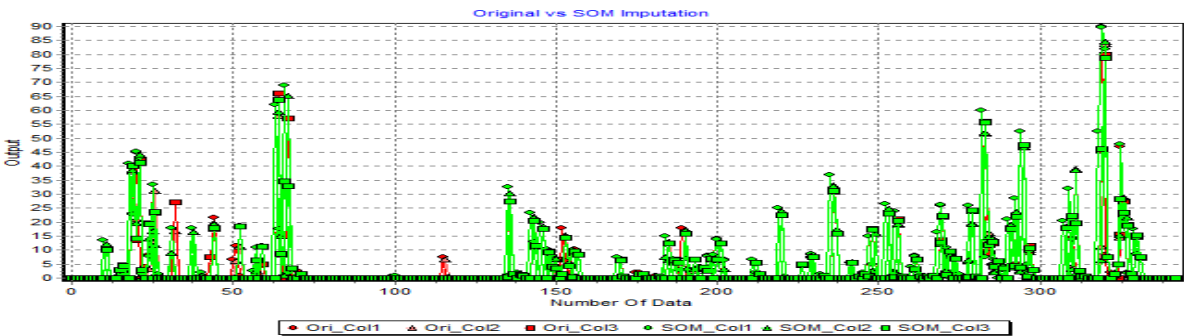


Fig. 3a. Comparison between original and imputed data using SOM Model for 40% days of hypothetical missing data set

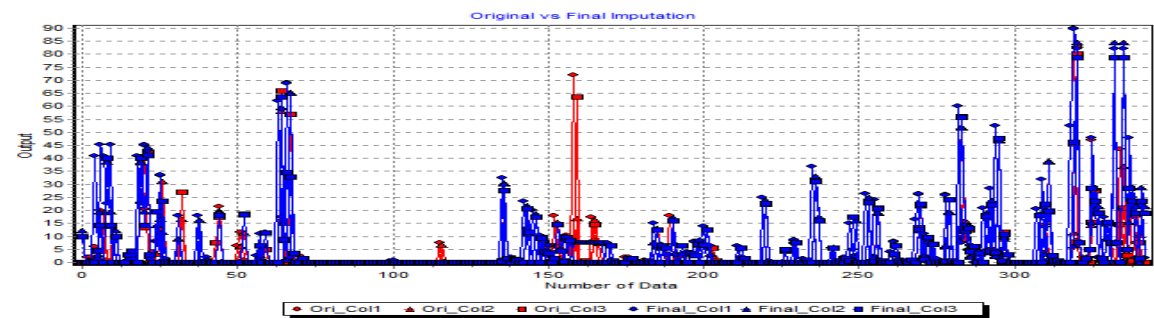


Fig. 3b. Comparison between original and imputed data using SOM-NNNeigh Hybrid Model for 40% days of hypothetical missing data set

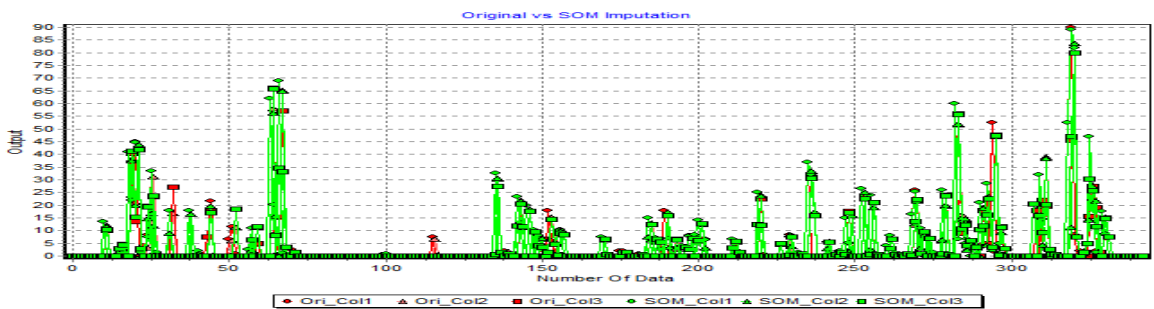


Fig. 4a. Comparison between original and imputed data using SOM Model for 60% days of hypothetical missing data set

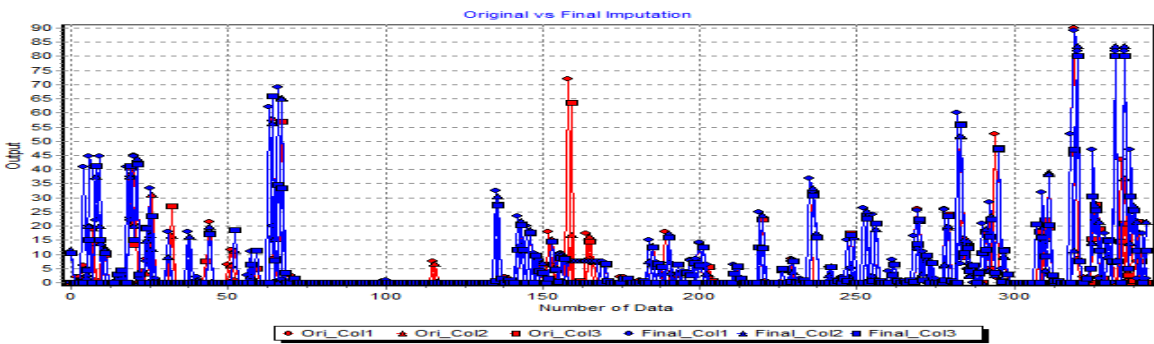


Fig. 4b. Comparison between original and imputed data using SOM-NNNeigh Hybrid Model for 60% days of hypothetical missing data set

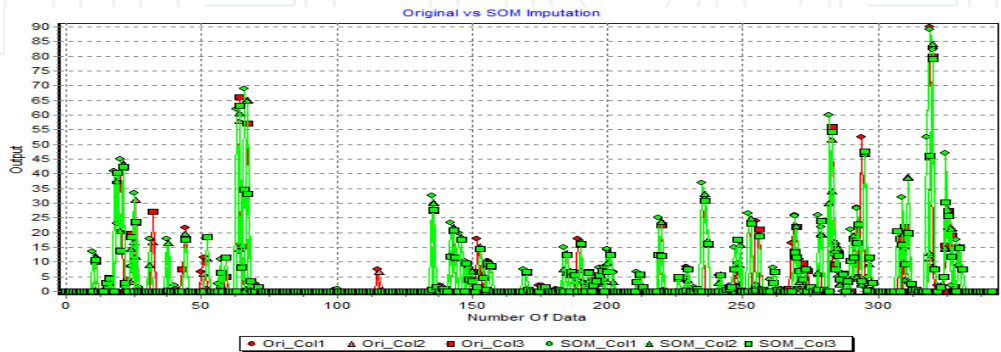


Fig. 5a. Comparison between original and imputed data using SOM Model for 70% days of hypothetical missing data set



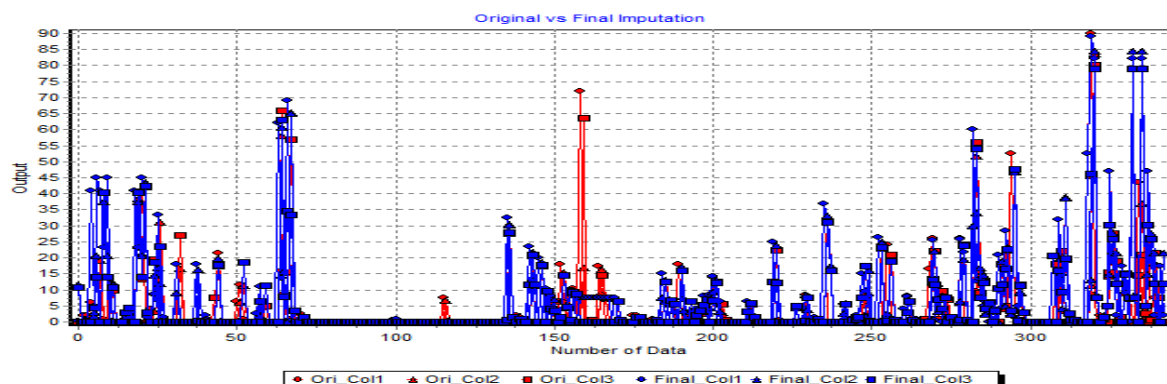


Fig. 5b. Comparison between original and imputed data using SOM-NNeigh Hybrid Model for 70% days of hypothetical missing data set

Based on Figure 2a to 5b, below are the general characteristics found at all rainfall stations in Perlis, Selangor and Johor at 20% to 70% of missing data:

- The imputed values are found to closely follow pattern of the original data at all percentages of missing data
- Imputed peak values produced by hybrid of SOM Model with NNeigh that diverge from the observed peaks are illustrated at all percentages of missing data.
- Peaks of imputed data produced by SOM-NNeigh Hybrid Model that diverge from the original data, did not occur at all in the SOM Model. This is because when all three recordings fail to produce any data, SOM Model ignore the missing data for that particular day.
- For up to 70% days of missing entries, most peaks of the imputed data coincide with the peaks of the original data.
- Even at the amount of 70% of missing data, pattern of imputed data fits almost perfectly that of the original data.

**80% of the days with missing data at rainfall station 6401002- Perlis**

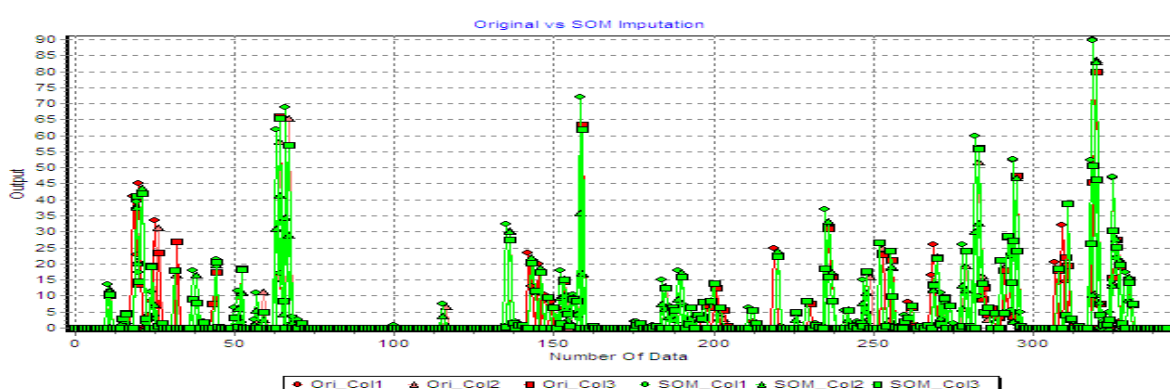


Fig. 6a. Comparison between original and imputed data using SOM Model for 70% missing up to two methods - 10% missing from all methods.



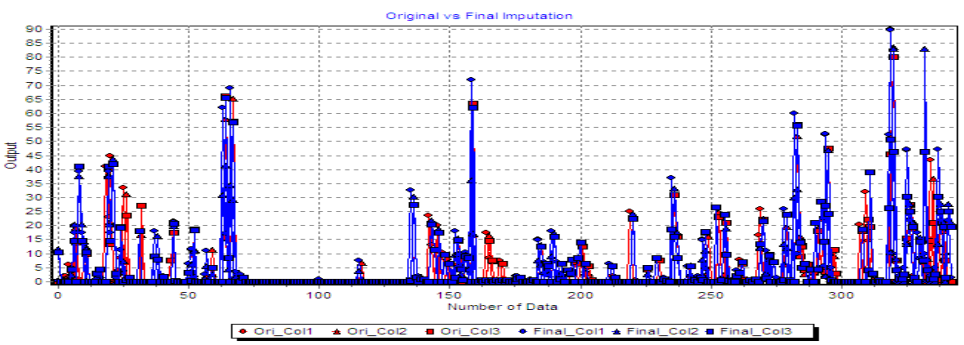


Fig. 6b. Comparison between original and imputed data using SOM-NNNeigh Hybrid Model for 70% missing up to two methods - 10% missing from all methods

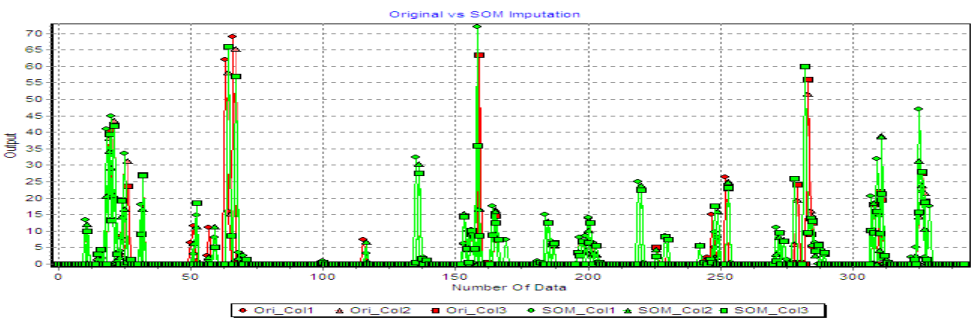


Fig. 7a. Comparison between original data and imputed data using SOM Model for 40% missing up to two methods - 40% missing from all methods

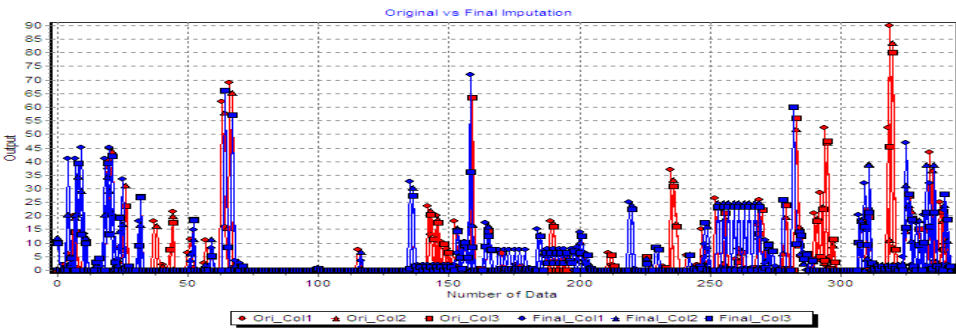


Fig. 7b. Comparison between original data and imputed data using SOM-NNNeigh Hybrid Model for 40% missing up to two methods - 40% missing from all methods

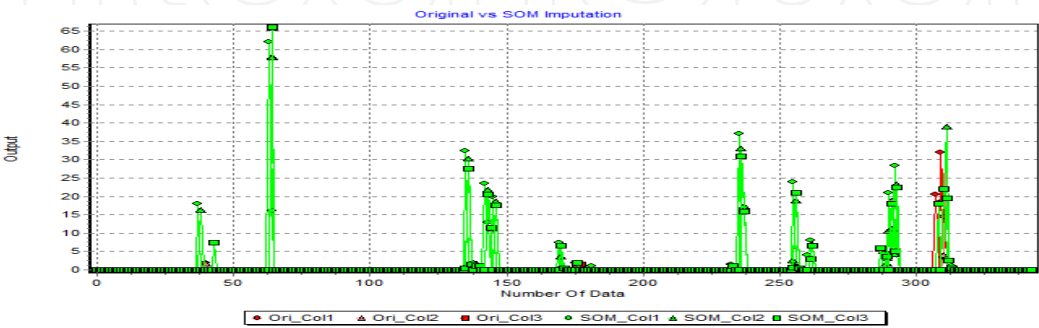


Fig. 8a. Comparison between original data and imputed data using SOM Model for 10% missing up to two methods - 70% missing from all methods

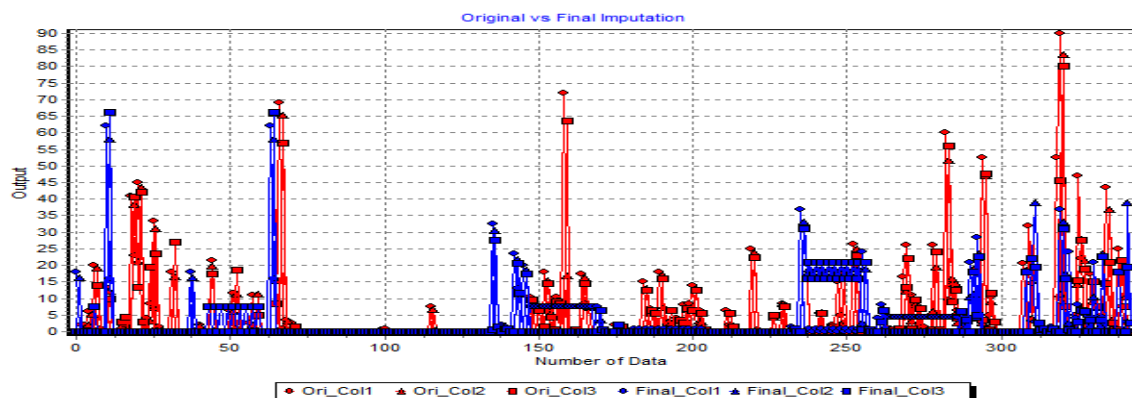


Fig. 8b. Comparison between original data and imputed data using SOM-NNeigh Hybrid Model for 10% missing up to two methods - 70% missing from all methods

Based on Figure 6a to 8b, findings on graphical presentation for the case of 80% days with missing data are as listed below:

- All rainfall stations in Perlis, Selangor and Johor exhibits similar results when SOM Model is applied. It can be seen that pattern of imputed rainfall data closely resembles that of the original values where peaks of the original data are mostly attainable. This applies where at least one entry of data is available in a particular day.
- When SOM Model is hybrid with NNeigh, the graphical presentation shows that in all cases the difference between predicted or imputed values and the observed or original data are more pronounced for 70% missing from all three methods, than for 10% missing from all three methods. As NNeigh caters for the case when data are not available from all three methods, it is rational to find that this model is reliable with lesser amounts of missing data from all three methods of recordings.

#### 4.4 Visualization Analysis on SOM Model

Self Organizing Method (SOM) used in Unsupervised ANN is commonly used as visualization aids. Relationships between vast amounts of data can be interpreted using this method. Data with similar characteristics end-up clustered together. Figure 9 to 15 shows the mappings of rainfall data distributions collected from different methods of recording at rainfall station 6401002 in Perlis for 20% to 80% of days with missing data.

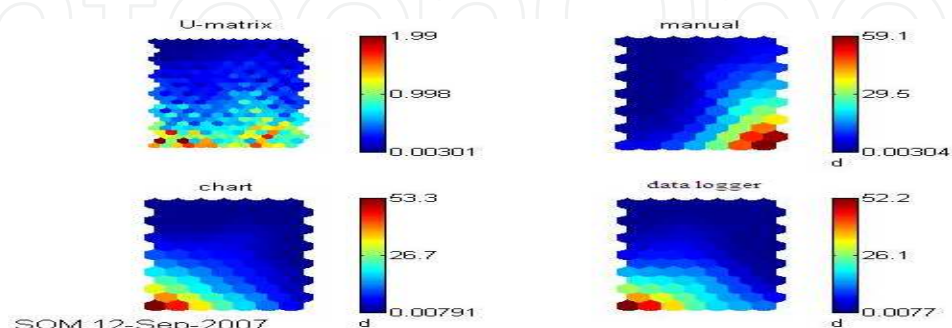


Fig. 9. SOM map for 20% days of hypothetical missing data set

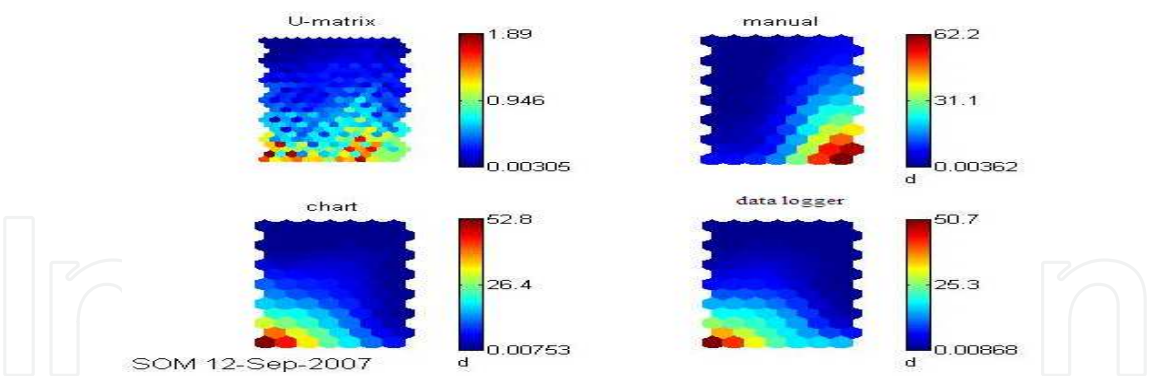


Fig. 10. SOM map for 40% days of hypothetical missing data set

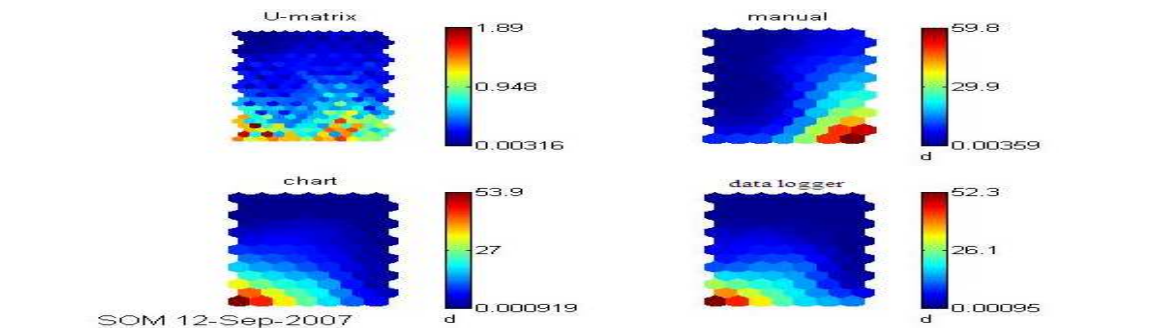


Fig. 11. SOM map for 60% days of hypothetical missing data set

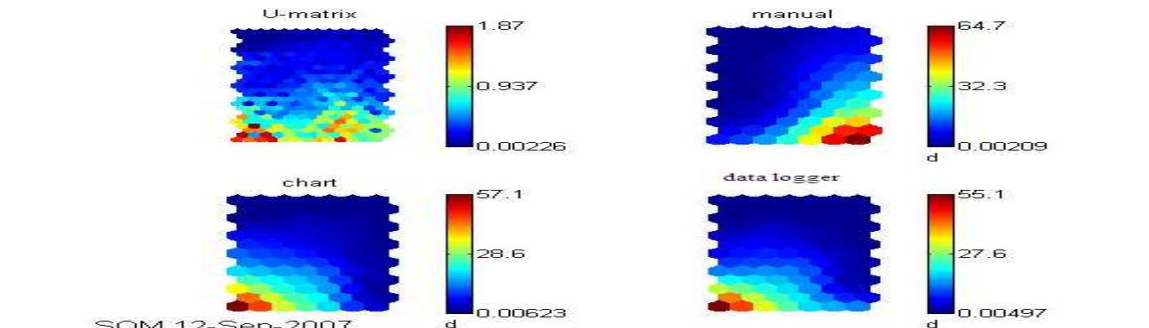


Fig. 12. SOM map for 70% days of hypothetical missing data set

Visualization Analysis on SOM for 80% days with missing data

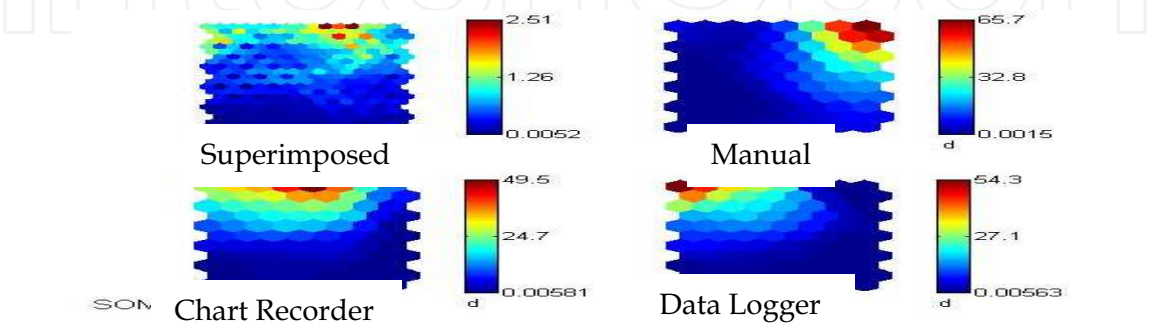


Fig. 13. Clustering of data at 70% missing up to two methods+ 10% missing from all three methods

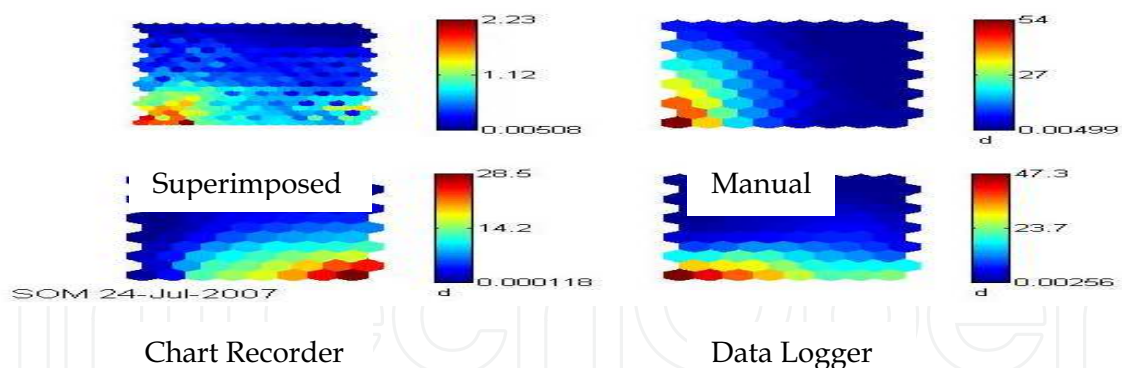


Fig. 14. Clustering of data at 40% missing up to two methods+ 40% missing from all three methods

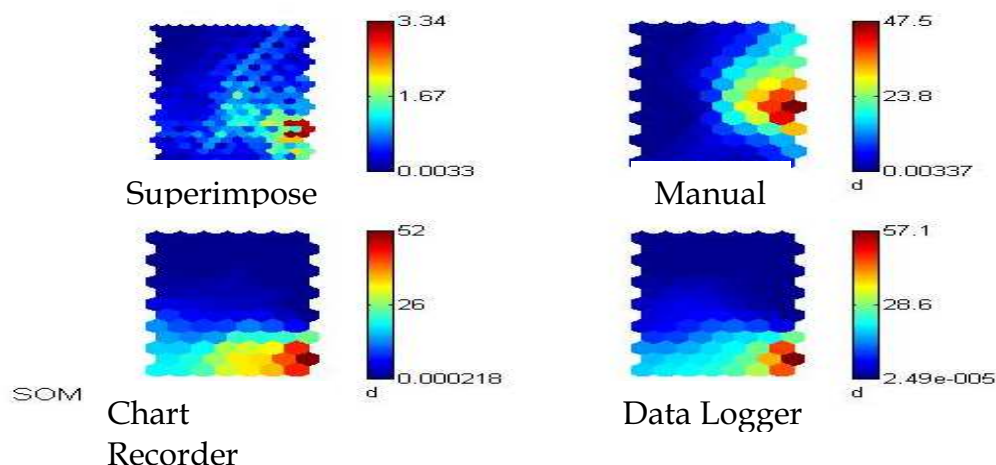


Fig. 15. Clustering of data at 10% missing up to two methods+ 70% missing from all three methods

In the SOM maps, wet days with higher values of rainfall depths are represented in red. Whilst, the darker blue represents dry days with approximately no rain. The colour classification visualizes the cluster structure of the maps according to rainfall depths in mm. Values on the colour bar are denormalized and show the original value range. The hexagonal grid is a unified distance matrix, commonly known as U-matrix. Each hexagon represents a node in the SOM. The U-matrix on the top left, shows superimposed image of rainfall data distribution obtained from all three methods of recordings.

## 5. Conclusions

This study has drawn several key conclusions as listed below:

- This study has successfully proven that the proposed models are feasible and can be reliably applied to emulate and substitute historical missing daily rainfall data when one, two or all three of the recordings are not available.
- The proposed SOM Model have proven to be reliable and accurate, even when they are applied to records with a maximum of 80% of days with missing data of up to two recording methods.

- c) The proposed hybrid of SOM-NNeigh models are both capable of yielding acceptable accuracy, even where a maximum of 40% of days are without any entries from all three recording methods.

## 6. Recommendation

It is recommended that SOM Model is used for cases where the occurrence of missing data is limited up to two methods of recordings, while SOM-NNeigh Hybrid Model is used when the occurrence of missing data from all three recording methods exist.

The advantages of using SOM Model and SOM-NNeigh Hybrid Model are its simplicity, ease of use, the fact that it does not require any distribution of input data and produce consistently accurate imputed values regardless of input data quality level. The fact that this model is able to produce reliable results at multiple rainfall stations in Perlis, Selangor and Johor, suggests that this model can be applied to any rainfall stations within Peninsular Malaysia encountering the same pattern of missing data. In short, a stable and generalized model has been achieved.

Due to the versatility and robustness of SOM and SOM-NNeigh Hybrid Models, it is recommended that this study is to be continued for future research works by incorporating other types of time series data with more than three variables, such as wind, temperature, humidity, etc.

## Acknowledgment

This work was supported by the Ministry of Science Technology and Innovation, Malaysia under Grant 04-02-03-SF0022. The author wish to thank Jabatan Pengairan dan Saliran (JPS), Malaysia for access to data.

## 7. References

- Aitken, A. P. (1974). Assessing Systematic Errors in Rainfall-Runoff Models. *Journal of Hydrology*. 20, 131-136.
- Porporato, A. and Ridolfi, L., (1997). Nonlinear analysis of river flow time sequences. *Water Resour. Res.*, 33(6), 1353-1367.
- Sani S. (1986). Rainfall Patterns In And Around Kuala Lumpur. Geography Department, Universiti Kebangsaan Malaysia. Malaysia.
- Sivakumar, B., Liong, S., Liaw, C. and Phoon, K., (1999). Singapore rainfall behaviour: chaotic?. *J. Hydrol. Eng., ASCE*, 4(1), 38-48.
- Tang, W.Y., Kassim, A. H. M. and Abu Bakar, S. H. (1996). Comparative studies of various missing data treatment methods - Malaysian experience *Atmospheric Research*, Volume 42, Issues 1-4, October 1996, 247-262





## **New Trends in Technologies**

Edited by Blandna ramov

ISBN 978-953-7619-62-6

Hard cover, 242 pages

**Publisher** InTech

**Published online** 01, January, 2010

**Published in print edition** January, 2010

This book provides an overview of subjects in various fields of life. Authors solve current topics that present high methodical level. This book consists of 13 chapters and collects original and innovative research studies.

### **How to reference**

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Marlinda A. Malek, Siti Mariyam Shamsuddin and Sobri Harun (2010). Restoration of Hydrological Data in the Presence of Missing Data via Kohonen Self Organizing Maps, New Trends in Technologies, Blandna ramov (Ed.), ISBN: 978-953-7619-62-6, InTech, Available from: <http://www.intechopen.com/books/new-trends-in-technologies/restoration-of-hydrological-data-in-the-presence-of-missing-data-via-kohonen-self-organizing-maps>

**INTECH**  
open science | open minds

### **InTech Europe**

University Campus STeP Ri  
Slavka Krautzeka 83/A  
51000 Rijeka, Croatia  
Phone: +385 (51) 770 447  
Fax: +385 (51) 686 166  
[www.intechopen.com](http://www.intechopen.com)

### **InTech China**

Unit 405, Office Block, Hotel Equatorial Shanghai  
No.65, Yan An Road (West), Shanghai, 200040, China  
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元  
Phone: +86-21-62489820  
Fax: +86-21-62489821



© 2010 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the [Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License](https://creativecommons.org/licenses/by-nc-sa/3.0/), which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen