Application of Self-Organizing Map (SOM) in Missing Daily Rainfall Data in Malaysia

Ho Ming Kang
Department of Mathematics, Faculty of Science,
Universiti Teknologi Malaysia,
81310 Skudai, Johor, Malaysia

Fadhilah Yusof
Department of Mathematics, Faculty of Science,
Universiti Teknologi Malaysia,
81310 Skudai, Johor, Malaysia

ABSTRACT
Self-organizing map (SOM) is applied to deal with missing daily rainfall data with different rainfall patterns in Peninsular Malaysia. In this study, stations from Damansara and Kelantan are focused and aimed to evaluate the effectiveness of SOM in clustering and imputation of missing data. The missing data that are imputed by SOM are evaluated by computing the mean square error (MSE) and coefficient correlation(R). Besides, the effects of the imputed data to the mean and variance of the rainfall data is also been observed. The clustering analysis showed that all the stations in Damansara are grouped distinctively, and having a good and even distribution of rain intensity as compared to Kelantan. Meanwhile it is also found that SOM is an excellent tool in estimation of missing data.

Keywords
Self-organizing map (SOM); missing values

1. INTRODUCTION
Missing data is a repeated problem encountered by most of studies especially in hydrology. Data are missing due to erroneous when sampling, insufficient of samples obtained, or problem in recording. Initially, the hydrologists usually used the methods of nearest neighbour weighting and inverse distance weighting method (IDWM) to impute the missing rainfall. However, the results reached the accuracy if there have a significant impact of neighbouring stations to the target station and when the data size is small (Teegavarapu and Chandramouli, 2005).

Self-organizing map (SOM) is a recent method that frequently applied in most of the hydrological studies. Kalteh and Hjorth (2009) used SOM to construct a complete database of runoff prediction in Northern Iran. Multiple imputation, multivariate nearest neighbour, multilayer perceptron (MLP) and means of SOM are the imputation methods used to estimate the missing data. Meanwhile, Koikkalainen and Horppu (2007) employed mean imputation; simulated randomness and donor imputation after the data are trained by TS-SOM (Tree-Structured SOM). Furthermore, Teegavarapu and Chandramouli (2005) explored the effectiveness of SOM by comparing with other classical methods like inverse-distance weighting method (IDWM). While, study of Juininen et al. (2004) used univariate (linear, spline and nearest neighbour interpolation) and multivariate (regression-based imputation, MLP, self-organizing map (SOM)), and a hybrid model to the air quality data sets in the APPETISE database. In Malaysia, there are also some studies that used artificial neural network (ANN) as in Ismail et al., 2010; Bustami et al., 2007; Talib and Abu Hasan, 2007. In particular, Marlinda et al. (2008) demonstrated the used of SOM with nearest neighbour imputation (NNeigh) in filling the missing rainfall recording data in Peninsular Malaysia.

Recently, the impact of climatic change causes the varying of rainfall patterns in Peninsular Malaysia. The high frequency of flood due to monsoon and non-monsoon seasons will have significant influence to the existence of missing rainfall data. Therefore, this study will use the SOM algorithm in treating this missing data problem. Besides, the clustering analysis by SOM on the two river basins namely Damansara and Kelantan, Malaysia also will be examined. The results will be discussed in Section 4.

2. BASIC OF SELF-ORGANIZING MAP
The well-known type of SOM used is Kohonen network which maps the input vectors onto a discrete map of 1 or 2 dimensions. Generally, it is a topologically ordered network because the input vectors are close to each other in the map. Kohonen network consists of a grid of output units and N input units. The input pattern with their own weights is fed and connected to the output unit by a small random numbers. As iteration increases, the winning output node is obtained and this node is actually a weight vector that has the smallest Euclidean distance to the input pattern. In addition, the weights of every node in the neighbourhood of the winning node are updated and this will move each node in the neighbourhood closer to the input pattern. At the same time, the learning rate and the radius of the neighbourhood will decrease as time increases. Finally if the parameters are well defined, the network will able to capture the clusters of the input data (Marlinda, 2008).

Before applying SOM on the rainfall data in Peninsular Malaysia, it is needed to process the data in order to fit the nature of SOM system. Normalization is important to ensure none of the variables used overwhelm the training result. The algorithm is then started by the initialization of random weights to each node. Batch algorithm is used in the training since it will process the whole data set to the map iteratively. Next, the SOM will search the best matching unit (BMU) by computing the Euclidean distance to the input vector.

$$BMU = \arg\min_i \|x - W_i\| \qquad (2)$$

where W is weight vector. The determination of the node that are within the BMU`s neighbourhood is made by calculating the radius of the neighbourhood, $\sigma(t)$ and it will decrease over time as

$$\sigma(t) = \sigma_0 \exp\left(-\frac{t}{\lambda}\right) \qquad (3)$$

where σ_0 is initial radius. As time increases, the weight vector of every node within the BMU`s neighbourhood are also needed to be adjusted by,

$$W(t+1) = W(t) + \eta(t)(V(t) - W(t)) \quad (4)$$

$$L(t) = L_0 \exp\left(-\frac{t}{\lambda}\right) \quad (5)$$

$$\eta(t) = \exp\left(-\frac{r^2}{2\sigma^2(t)}\right), t = 1,2,3,\ldots \quad (6)$$

where r is the distance of a node from the BMU, $W(t)$ is the weight, $L(t)$ is the learning rate, $\eta(t)$ is the amount of influence a node`s distance from the BMU has on its learning, and $V(t)$ is the input vector . The effect of learning rate will be lowered as it is proportional to the distance of the node from the BMU. The process will be trained until maximum iteration is reached.

In this study, Neural Network toolbox from MATLAB 2007 is developed accordingly. The rainfall data from Damansara and Kelantan are separated into testing data with no missing data and validation data which comprises of some missing data respectively. The training data will be used in the SOM training process meanwhile the validation data is used to evaluate the performances of SOM. By trial and error, the most suitable map size is 25x25.

3. STUDY AREA

Daily rainfall data of the 23 stations are obtained from the Department of Irrigation and Drainage (DID) Malaysia. Table 1 describes the stations with their geographical coordinates and the percentage of missing values from Damansara and Kelantan river basins respectively.

Kelantan river basins are located in the eastern part of Peninsular Malaysia with annual mean rainfall from 2,032mm to 2,540mm. North-east monsoon is the main factor that contributes to heavy rainfall to Kelantan that occurs from November to February. Meanwhile for Damansara river basin, the most important urban centre is situated here. The rapid development in the last 50 years caused the agriculture lands to be converted into townships which consist of commercial and industrial areas. Heavy rainfall normally happened in Damansara during the inter-monsoon season which usually occurs from March to April and from September to October.

Table 1: The geographical coordinates of the stations with missing values in Damansara, Johor and Kelantan.

Code	Name of Station	Latitude	Longitude	% of Missing Values
Damansara				
1.	Sek.Men Sri Aman	3.1032	101.6294	6.763
2.	Sek Men Keb Taman Sea	3.1106	101.6174	2.820
3.	Sek Ren China Yuk Chai	3.1151	101.6077	6.846
4.	Sek Ren Dsara Utama	3.1490	101.6244	8.708
5.	Tropicana Golf	3.1369	101.5921	5.969
6.	Masjid Sg Penchala	3.1627	101.6275	8.324
7.	Sek Men Dsara Jaya	3.1303	101.6179	9.310
8.	Balai Polis Sea Park	3.1222	101.6354	16.05
9.	Sek Men Sri Permata	3.1019	101.6141	12.40
10.	Kompleks Fas	3.1091	101.6355	10.10
Kelantan				
11.	Brook	4.6764	101.4844	3.505
12.	Blau	4.7667	101.7569	1.205
13.	Gunung Gagau	4.7569	102.6556	5.367
14.	Hau	4.8167	101.5333	7.284
15.	Gua Musang	4.8847	101.9694	1.889
16.	Chabai	5.0000	101.5792	5.887
17.	Kg. Aring	4.9375	102.3528	3.395
18.	Gemala	5.0986	101.7625	0.575
19.	Balai Polis Bertam	5.1458	102.0486	6.681
20.	Dabong	5.3778	102.0153	7.667
21.	Kg. Laloh	5.3083	102.2750	9.611
22.	Ulu Sekor	5.5639	102.0083	3.724

4. RESULTS

The quality and the performance of SOM to the missing rainfall data in Damansara and Kelantan are evaluated by using visualization analysis and by computing the mean square error (MSE) and correlation coefficient (R).

4.1 Clustering Analysis

Self-organizing map (SOM) is always used to visualize the relationship between the data given and clustered the data with similar characteristics. The blue colour in the maps represents the dry days or no rain, meanwhile the red colour represents the high rainfall intensity or heavy rain. In other words, the cluster of similar characteristics and properties will be visualized by the colours. In the process of SOM, it will generate a matrix where each entry is the Euclidean distance (in input space) between neighbouring representatives, and known as U-matrix (Izenman, 2008). It is also an image of the combination of data distribution obtained from the variables used in the analysis (Marlinda, 2008).

Figure 1 shows the component planes of the stations in Kelantan. The upper-left of the component planes is the U-matrix and the "components" are actually referred as the stations in Kelantan areas. The colour maps show that all the stations in Kelantan are exhibiting different rainfall patterns from each other. Each neighbourhood shows a significance difference in terms of rainfall amount. It also indicates that the division of homogeneous stations is hard and ambiguous. This phenomenon appeared because Kelantan is an area which is separated from other states by Titiwangsa Range and the existence of second highest peak in Peninsular Malaysia – Gunung Korbu. Therefore most of the stations are clearly separated and sheltered by the mountain, and hence the distribution of rainfall becomes unequal and sparsely distributed. Therefore the estimation of the missing values in Kelantan which is based on neighbouring stations may not be accurate and can lead to biased results.

Different from Kelantan, Damansara shows distinct clusters among the stations. In Figure 2, the component planes are arranged into three groups that correspond to three spectral bands. The component planes show that the stations from Sri Aman to Sea Park, Yuk Chai, Dsara Utama, Tropicana Golf and Dsara Jaya (1 group) are different substantially between Taman Sea (1 group) and Sg. Penchala (1 group). Within each group, there are some differences between each station but it is not significant when compared to each spectral bands. In short, the clusters are actually considered as fairly separated and distributed. This is because Damansara is an urban area with the most densely populated in Peninsular Malaysia, and

hence the distribution of rainfall varies slightly due to the absence of intervention of high lands and mountains.

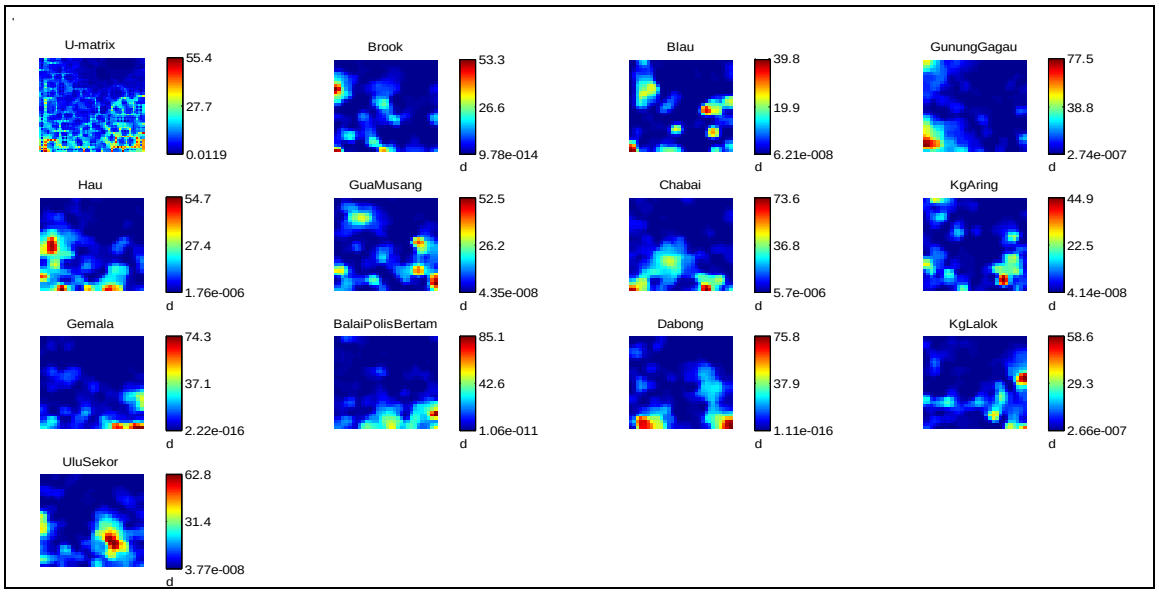

Figure 1: Component Planes of Kelantan areas

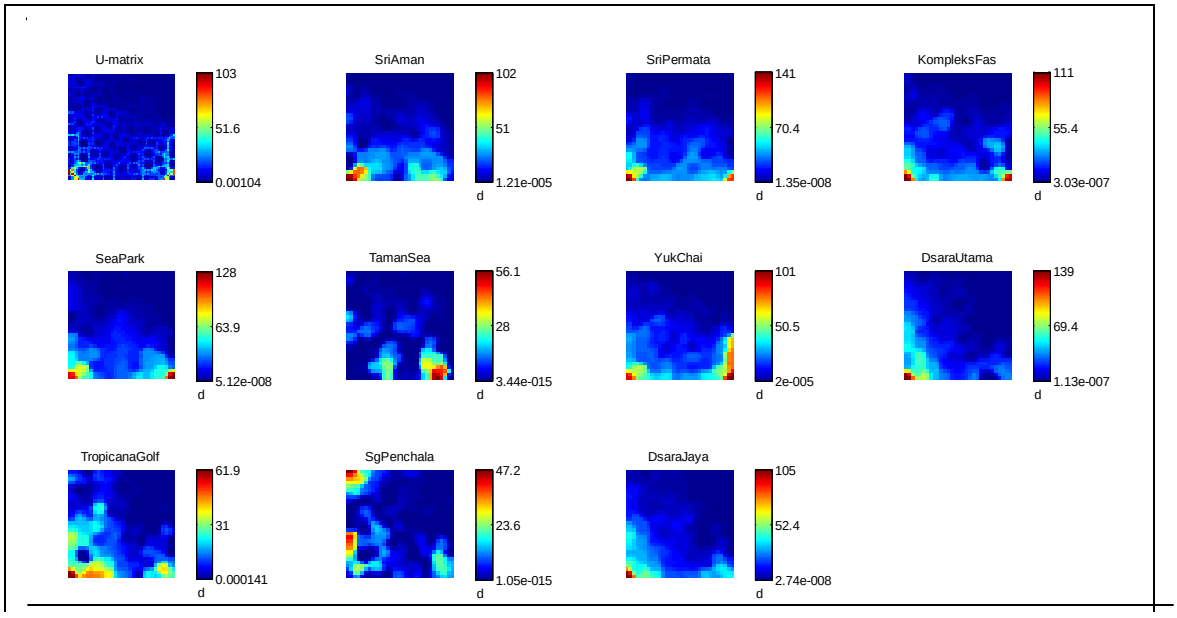

Figure 2: Component Planes of Damansara areas

4.2 Accuracy Of Som

Table 2: Quantization and Topographic Error

Study Area	Quantization Error	Topographic Error
Kelantan	3.7699	0.0097
Damansara	0.7766	0.0188

Table 2 shows the quantization error and topographic error. Topographic error used to estimate the complexity of the output space (Marlinda, 2008) meanwhile quantization error estimated the closeness of the best matching unit (BMU) to the input data. From the results obtained, Damansara is having small quantization error when compared to Kelantan. It is also shows that the BMUs produced in Damansara are the optimum and close to the input data with minimum error.

Table 3: The Results of MSE and R to the Stations in Damansara and Kelantan

Code	Name of Station	MSE	R
Damansara			
1.	Sri Aman	3.3672	0.8531
2.	Sri Permata	2.9386	0.8842
3.	Kompleks Fas	2.8512	0.8725
4.	Sea Park	3.2538	0.8806
5.	Taman Sea	3.2108	0.8338
6.	Yuk Chai	2.9190	0.8962
7.	Dsara Utama	4.0102	0.7674
8.	Tropicana Golf	3.0460	0.8603
9.	Masjid Sg Penchala	3.2741	0.8803
10.	Dsara Jaya	4.0399	0.8083
Kelantan			
11.	Brook	3.6412	0.7559
12.	Blau	4.2416	0.6625
13.	Gunung Gagau	4.5745	0.8448
14.	Hau	4.4267	0.6578
15.	Gua Musang	4.7285	0.6194
16.	Chabai	3.7553	0.7007
17.	Kg. Aring	5.6079	0.5191
18.	Gemala	4.1040	0.4824
19.	Balai Polis Bertam	3.7959	0.6942
20.	Dabong	4.2499	0.7467
21.	Kg. Laloh	5.6055	0.4366
22.	Ulu Sekor	5.6541	0.6643

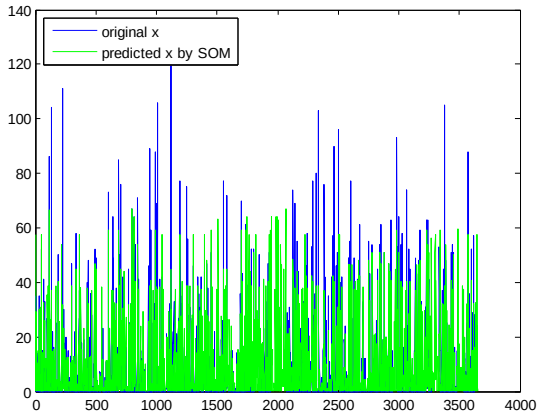

Figure 3: Comparison of Estimated Values by SOM and Observed Data in Komplek Fas, Damansara

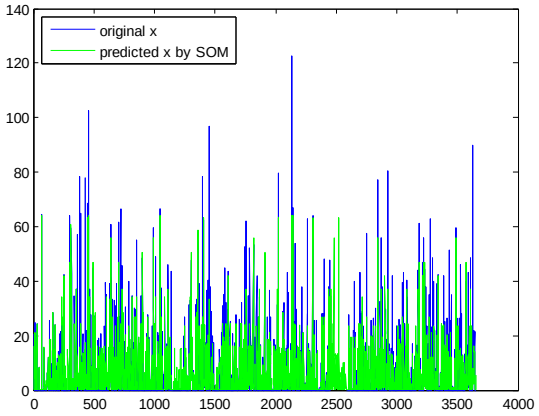

Figure 4: Comparison of Estimated Values by SOM and Observed Data in Brook, Kelantan

Table 3 exhibits the results of mean square error (MSE) and correlation coefficient (R) of the SOM imputation. The performance of SOM in estimating the missing data in Damansara has show encouraging results. In particular, the errors produced for stations in Damansara are lower and the correlation coefficients (R) are also comparatively higher when compared to stations in Kelantan. This is because the estimation of missing data in Kelantan is quite complicated due to the topological influence and the impacts of distance and elevation. That has consequently results in higher MSE and R produced by SOM. Figure 3 and 4 illustrate the comparisons between the estimated values from SOM and observed rainfall data in Kompleks Fas, Damansara and Brook, Kelantan respectively. Based on the graphs obtained, it is verified that SOM is able to predict the data accurately even when the rainfall amount is at the maximum.

Table 4: Confidence Interval of Mean and Variances in Damansara and Kelantan

	Confidence Interval	
	Mean	Variance
Damansara		
Sri Aman	[7.4580,8.5308]	[261.5833,286.7333]
Sri Permata	[6.8088,7.8178]	[231.3705,253.6156]
Kompleks Fas	[6.5541,7.4735]	[192.0969,210.5661]
Sea Park	[7.5073,8.6773]	[311.1966,341.1167]
Taman Sea	[6.0609,6.9895]	[195.9651,214.8062]
Yuk Chai	[7.0596,8.0222]	[210.6043,230.8528]
Dsara Utama	[7.8088,8.8622]	[252.1804,276.4263]
Tropicana Golf	[6.1509,7.1001]	[205.0168,224.7281]
Masjid Sg Penchala	[7.3817,8.5053]	[238.1926,261.0937]
Dsara Jaya	[5.6215,6.4769]	[166.3346,182.3269]
Kelantan		
Brook	[5.8609,6.6041]	[125.5369,137.6066]
Blau	[5.2463,6.0901]	[161.8146,177.3723]
Gunung Gagau	[9.1636,10.6184]	[481.0836,527.3374]

Hau	[5.7942,6.5614]	[133.7864,146.6493]
Gua Musang	[5.5826,6.4312]	[163.6733,179.4097]
Chabai	[4.8158,5.5932]	[137.3980,150.6082]
Kg. Aring	[6.5220,7.5734]	[251.2403,275.3958]
Gemala	[4.1255,4.9445]	[152.4994,167.1615]
Balai Polis	[4.8783,5.6854]	[148.0762,162.3130]
Bertam Dabong	[6.5830,7.6248]	[246.7257,270.4472]
Kg. Laloh	[6.4580,7.5144]	[253.6484,278.0355]
Ulu Sekor	[7.1496,8.3390]	[321.5974,352.5174]

Table 5: Descriptive Statistics between Original Data and Imputed Data in Kelantan and Damansara

Damansara	Imputed Data		Descriptive Statistics of Data (Ignoring Missing Data)	
	Mean	Variance	Mean	Variance
Sri Aman	8.0507	266.8682	7.9944	273.5815
Sri Permata	7.4075	241.8181	7.3133	241.9829
Kompleks Fas	7.1565	204.2527	7.0138	200.9079
Sea Park	8.2363	323.6797	8.0923	325.4705
Taman Sea	6.4892	201.3703	6.5252	204.9536
Yuk Chai	7.4497	212.4219	7.5409	220.2642
Dsara Utama	8.1603	257.2816	8.3355	263.7473
Tropicana Golf	6.6629	210.8856	6.6258	214.4204
Masjid Sg Penchala	7.7812	241.2586	7.8935	249.1179
Dsara Jaya	5.7590	163.9885	6.0492	173.964

Kelantan	Imputed Data		Descriptive Statistics of Data (Ignoring Missing Data)	
	Mean	Variance	Mean	Variance
Brook	6.2127	129.7184	6.2325	131.2949
Blau	5.6816	168.3428	5.6682	169.2367
Gunung Gagau	9.8944	493.4129	9.8910	503.1498
Hau	6.0670	134.936	6.1778	139.9229
Gua Musang	5.9512	168.953	6.0069	171.1806
Chabai	5.1296	139.2707	5.2045	143.7002
Kg. Aring	6.8975	255.5106	7.0477	262.7641
Gemala	4.5208	158.7021	4.5350	159.4942
Balai Polis	5.1177	147.3505	5.2819	154.8681
Bertam Dabong	6.9951	248.5352	7.1039	258.0425
Kg. Laloh	6.5887	243.0013	6.9862	265.2827
Ulu Sekor	7.5467	325.4596	7.7443	336.3483

Table 4 summarises the confidence interval of mean and variance while Table 5 describes the descriptive statistics of Kelantan and Damansara after the imputation of SOM and without imputation. The evaluation is carried out by observing the mean and variance of the imputed data to the confidence interval of the rainfall data with missingness. The results show that SOM is working efficiently and effectively in dealing with missing rainfall data in those two river basins. All of the mean and variance from the imputation data by SOM for Damansara and Kelantan lie within the confidence interval of the mean and variance of the data without imputation. Therefore, we may conclude that SOM is a good tool in clustering and estimating the missing rainfall data with different patterns and characteristics.

5. CONCLUSION

The SOM can be reliably applied in most of the hydrological study, especially in dealing with missing values on rainfall data. In this study, daily rainfall data of Damansara and Kelantan are used. Self-organizing map (SOM) is then applied to identify the relationship between stations, to cluster stations with similar characteristics, and to impute the missing rainfall data. The results showed that the patterns and distributions of Kelantan and Damansara are different from each other due to the geographical locations of the stations and the effects of monsoon. Damansara is the only area which has a fairly and evenly distribution of rainfall amount. Besides, the clustering in Damansara is also distinct and therefore the application of SOM in dealing with missing data becomes effective and efficient. The quantization and topographic errors of SOM is found to be at the minimum when applied to Damansara data.

All stations of different areas are compared by computing the mean square error (MSE) and correlation coefficient (R). Besides, the descriptive statistics after the imputation of SOM are also compared. The results indicated SOM is a good estimator as the errors produced is low, especially applied to data in Damansara. Moreover, the mean and variance produced by the data after SOM imputation are also located within the confidence interval of the mean and variance from the data without imputation. Therefore, it has a strong basis to prove that SOM imputation on missing rainfall data are very strong and highly reliable, even in an uneven distribution of rainfall amount as in Kelantan.

6. ACKNOWLEDGEMENT

We thank to the fund and support from Zamalah Scholarship, Universiti Teknologi Malaysia (UTM).

7. REFERENCES

[1] Teegavarapu, R. S. V. and Chandramouli, V. 2005. Improved Weighting Methods, Deterministic and Stochastic Data Driven Models for Estimation of Missing Precipitation Records. *Journal of Hydrology*, 312, 191-206.

[2] Kalteh, A. M. and Hjorth, P. 2009. Imputation of Missing Values in a Precipitation Runoff ProcessDatabase. Hydrology Research, 40.4, 420-432.

[3] Koikkalainen, P. and Horppu, I. 2007. Handling Missing Data with the Tree-Structured Self-Organizing Map.

Proceedings of International Joint Conference on Neural Network, Orlando, Florida USA, August 12-17.

[4] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M. 2004. Methods for Imputation of Missing Values in Air Quality Data Sets. Atmospheric Environment 38, 2895-2907.

[5] Ismail. S., Samsudin, R. and Shabri, A. 2010. River Flow Forecasting: A Hybrid Model of Self-Organizing Map and Least Square Support Vector Machine. Hydrol. Earth Syst. Sci. Discuss. 7, 8179-8212.

[6] Bustami, R., Bessaih, N., Bong, C. and Suhaila, S. 2007. Artificial Neural Network for Precipitation and Water Level Prediction of Bedup River. IAENG International Journal of Computer Science, 34:2.

[7] Talib, A. and Abu Hasan, Y. 2007. The Application of Artificial Neural Network for Forecasting Dam Spillage Events. Universiti Sains Malaysia, Penang Malaysia.

[8] Marlinda A. M, S. Harun, S. M. Shamsuddin and I. Mohamad 2008. Imputation of Time Series Data via Kohonen Self-Organizing Maps in the Presence of Missing Data. World Academy of Science, Engineering and Technology, 41, 501-506.

[9] Izenman, A. J. 2008. Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning. Springer: New York.