An unsupervised machine learning approach for estimating missing daily rainfall data in peninsular malaysia

Wing Son Loh[1*], *Wei Lun* Tan[1], *Ren Jie* Chin[2], *Lloyd* Ling[2], *Sheong Wei* Phoon[1] and *Choon Sen* Seah[3]

[1]Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, 43000 Kajang, Malaysia
[2]Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, 43000 Kajang, Malaysia
[3]Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

Abstract. Rainfall data plays a vital role in various fields including agriculture, hydrology, climatology, and water resource management. Stakeholders had raised concerns over the issue of missing rainfall data as it presents a huge obstacle in achieving reliable climate forecasts. Therefore, it is necessary to perform accurate estimation for the missing daily rainfall data. Each year, the peninsular Malaysia experiences a significant rainfall event during the monsoon period due to the North-East monsoon (NEM) wind. The intricate spatial rainfall dynamics requires a computational model, capable of generating accurate estimates and deciphering hidden patterns from the missing data. An unsupervised machine learning model known as the Self-Organising Feature Map (SOFM) is developed to estimate the missing daily rainfall across 10 rainfall stations during the NEM period between 2010 and 2020. The SOFM exhibited reliable performance across the percentage of missingness between 10% to 50%. Below 50% missingness, the evaluated statistical metrics, coefficient of determination (R^2) is attained above 0.5, ranging between 0.504 and 0.915; root mean square error (RMSE) between 15.9 to 22.7. The feature maps enabled the visualisation of the relationship between the rainfall intensity and studied rainfall stations. The feature maps suggested that the studied rainfall stations are inhomogeneous.

1 Introduction

Addressing missing values in the context of rainfall data remains an ongoing challenge in hydrological and environmental research studies, with data gaps arising from events such as inaccuracies in rainfall measurement techniques, rain gauge station relocations, and instrument malfunctions. Subsequently, the existence of daily missing rainfall data is known to be a huge obstacle in preserving a serially complete real-time rainfall database. Conservatively, financial losses amounted to $57 billion per year was estimated for

*Corresponding author: lohws@utar.edu.my

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

worldwide flood-related disasters which were reported between 1990 and 2020 [1, 2]. In particular, the weather in the Peninsular Malaysia is affected by the North-East monsoon (NEM) season each year where a significant rainfall event takes place. The NEM is associated with longer span of rainfall occurrences, with relatively larger rainfall amount. The Peninsular Malaysia is thus vulnerable to flooding due to the severe rainfalls during the NEM. In the recent decades, numerous flood-related disasters have occurred around Peninsular Malaysia during this period of time, resulting in substantial financial and economic losses, the destruction of crops, assets and buildings, as well as concerns about water quality deterioration for consumption which affects human health. In addition, there were also unfortunate reports of fatalities. On top of that, the research findings from the World Bank and the Asian Development Bank (ADB) in Asia indicated that due to the effects invited by climate changes, the median increase in the population impacted by an extreme river flood is projected to be around 102,290 people by the years 2035-2044 [3].

Precise rainfall data holds immense importance across various domains, encompassing hydrological modelling, water resource management, and climate research. Valid analyses and robust hydrological models could not be produced without a complete and accurate rainfall data. As the rainfall data contain gaps in the daily observations, relying on the available data especially with high proportion of missing values would introduce biases and hinder important information associated with the hydrological variables under consideration [4]. Hence, estimating missing daily rainfall amount under the highly complex synoptic rainfall dynamics within Peninsular Malaysia served as one of the main motivations in this study to generate a complete and reliable rainfall dataset.

Existing studies related to rainfall modelling have suggested that the estimation results under conventional techniques such as linear regression models, deletion method, mean imputation, and normal ratio method were less promising [5]. On the other hand, machine learning based models which are highly popular in the recent decades such as the support vector machine (SVM), decision tree (DT), k-nearest neighbourhood (kNN), and the artificial neural network (ANN) yielded better results especially for missing rainfall estimations [6, 7]. Additionally, the machine learning based models have an added advantage of being non-parametric which have no statistical assumption asserted on the data variable distributions itself. For instance, the inverse distance weighting (IDW) is a non-machine learning based model that dependents greatly on the homogeneity assumptions of the rainfall station studied [8]. Thus, the IDW mechanism operates with an underlying assumption that nearer station towards the target rainfall station are homogenous and holds higher weightage in influencing the rainfall dynamics when performing the estimations. Moreover, one of the major drawbacks of relying heavily on rainfall station homogeneity is that the daily rainfall observations from the stations classified as homogeneous are all missing at the same time. As a result, there are no observations that are available from other neighbouring rainfall stations which could comply with the station homogeneity assumption of the models.

Nevertheless, it is crucial to construct a reliable computational model that is capable of producing robust estimations for the missing daily rainfall data. In particular, this study involves developing an unsupervised machine learning based estimation model, known as the self-organising feature map (SOFM) model with the main objective of estimating missing daily rainfall data over the Peninsular Malaysia. The SOFM model was trained by utilising the spatial daily rainfall data collected from the different rainfall stations. The SOFM can be seen as a variation of the ANN where it consists of the Kohonen map in its network architecture. Unlike the design of ANN which has the hidden and output layers respectively, the SOFM conducts the training process by mapping the input instances to the Kohonen map directly. The Kohonen map act as a two-dimensional layer map that projects input patterns of high dimensionality onto the nodes which were arranged in a hexagonal-like structure within the map [1, 9]. Ingeniously, the Kohonen map in the SOFM model possesses the

ability to preserve topological properties within the input space for which the nodes in the Kohonen map were arranged based on the similarities in the learned instances towards the input feature patterns, captured by the weight vectors [10]. Together, the unsupervised learning mechanism of the SOFM model enables close density estimations and non-parametric projections for the input variables without preliminary identifications of the target label classes. Past cases of SOFM related applications in rainfall studies were discovered such as Ho and Yusof which predicted the dry and wet spells, and then estimated the daily rainfall values between 1996-2007. The estimation was performed based on two regional clusters in Peninsular Malaysia where the rainfall stations in the specified zone were a collection of neighbouring stations in the same zone. Under the proportion of missing values between 0.5% to 16%, the NSE values reported between 0.4366 and 0.8962 [11]. On another hand, Chin et al. applied the IDW and local polynomial interpolation (LPI) technique to compare the annual rainfall intensity. It was concluded that the IDW achieved a maximum NSE of 0.65 while the LPI achieved the best performance with NSE of 0.69 [8]. In another study, Angkool et al. discovered that the multiple linear regression model and ANN could provide estimation results of NSE value approximately 0.798 and 0.804 respectively [5].

2 Methodology

2.1 Study area and daily rainfall data

The Peninsular Malaysia contains 11 states and 2 federal territories with numerous river systems and mountain ranges. Disparate mountain elevations and unique geographical settings could be discovered suggesting the inhomogeneity in the rainfall characterisation. The geographical coordinates of the 15 studied rainfall stations were provided in Table 1. The map of Peninsular Malaysia with the corresponding location of the 15 studied rainfall stations is also illustrated in Figure 1.

Table 1. The geographic coordinates of the 15 studied rainfall stations.

Station Code	Station Name	Longitude	Latitude
48603	Alor Setar	100° 24'E	6° 12'N
48642	Batu Embun	102° 21'E	3° 58'N
48601	Bayan Lepas	100° 16'E	5° 18'N
48632	Cameron Highlands	101° 22'E	4° 28'N
48604	Chuping	100° 16'E	6° 29'N
48625	Ipoh	101° 06'E	4° 34'N
48672	Kluang	103° 19'E	2° 01'N
48615	Kota Bharu	102° 17'E	6° 10'N
48618	Kuala Terengganu	103° 06'E	5° 23'N
48657	Kuantan	103° 13'E	3° 47'N
48665	Melaka	102° 15'E	2° 16'N

48679	Senai	103° 40'E	1° 38'N
48620	Sitiawan	100° 42'E	4° 13'N
48647	Subang	101° 33'E	3° 07'N
48653	Temerloh	102° 23'E	3° 28'N

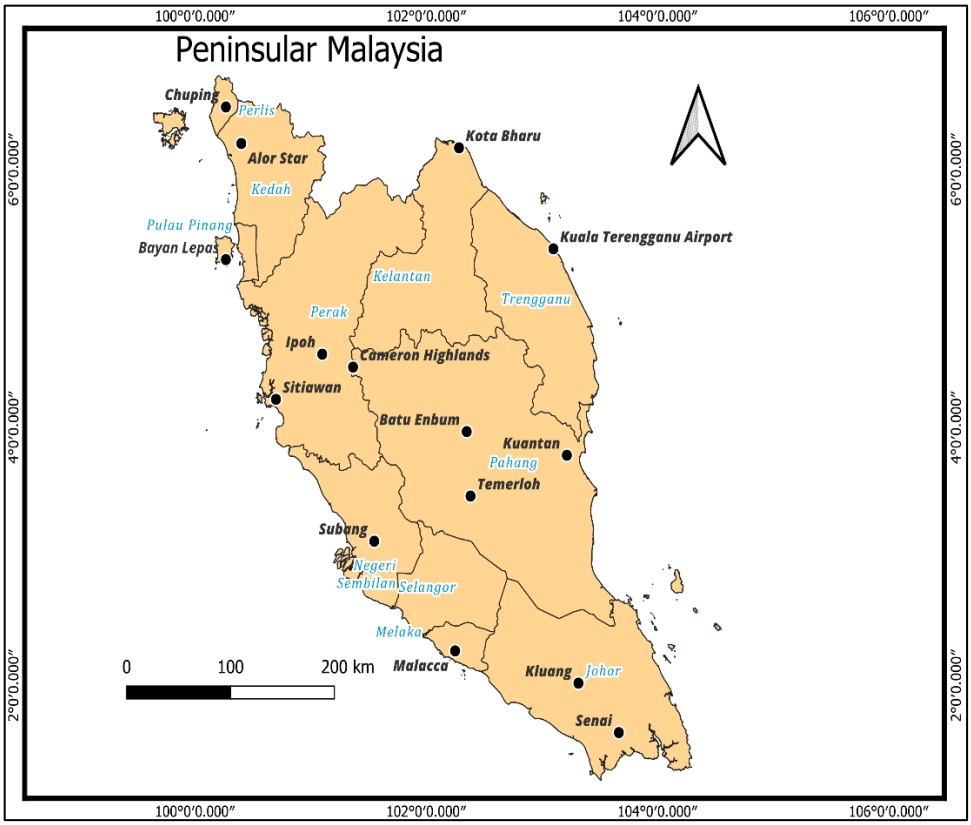

Fig. 1. Location of the 15 studied rainfall stations in the Peninsular Malaysia.

2.2 Missing Data Mechanism

In the context of missing data, the types of data missingness can be distinguished into three mechanisms, namely the missing not at random (MNAR), the missing completely at random (MCAR), and missing at random (MAR) [12]. Missing data classified as MNAR are assumed to be related to the missing data themselves. Hence, it is not appropriate to estimate the missing data using the complete observations that is available. On the other hand, missing data estimation techniques such as involving direct discarding the missing data values falls under the MCAR mechanism. According to the literatures, the MCAR is only appropriate for the case of small missing data percentages. On top of that, missing data under the MCAR mechanism are assumed to be independent of both complete and missing observations. Thirdly, the MAR mechanism states that the missingness of the data is independent of the values that are unobserved, given the information obtained from the observed data. The underlying assumption is that the missing data is considered as a random subset of the

observed data values. Hence, it is plausible that the estimation of missing daily rainfall data deals with the MAR mechanism due to the existence of statistical dependence between rainfall values.

In particular, the MAR mechanism contradicts with the homogeneous assumptions for rain gauge stations because the MAR suggests that missingness of the daily rainfall data is unrelated to the unobserved daily rainfall data values, but station homogeneity implies that the target station shares similar rainfall dynamics with the neighbouring homogeneous stations with an induced relationship dependency. Therefore, inhomogeneous rainfall stations were selected in this study, and the missing daily rainfall data were assigned by data amputation according to a stipulated missing percentage that is ranged between 10% to 50% under the MAR mechanism. In brief, the MAR mechanism in this study as shown in Equation (1) refers to the probability of a given daily rainfall observation, x coming from the database consisting of the missing data, x_{MD}, which was assumed to depend on an observable daily rainfall value from the complete data, x_{OBS}, but independent on the missing observation itself.

$$\Pr(x_{MD}|x) = \Pr(x_{MD}|x_{0BS}) \tag{1}$$

2.3 Missing Daily Rainfall Data Estimation

2.3.1 Self-Organising Feature Map (SOFM) Model

With growing interest in hydrological applications such as satellite imagery classifications [13], rainfall forecasts, runoff modelling and other related analyses, SOFM was commonly reported as a promising tool in visualising pattern projections because of its topologically ordered mapping within the output that clusters similar neurons in a lattice structure. The data distributions could be observed based on the illustrated Kohonen map as it preserves the topological information within the projection space. In this particular study, the SOFM model was first trained using the available complete daily rainfall observations. After the training phase ended, the Kohonen map storing all computed weight vectors is produced. The rainfall intensities and possible patterns / relationships arose from the studied rainfall stations can be examined from the individual feature maps for each variables. The feature maps were extracted from the trained Kohonen map based on the 15 studied rainfall stations.

2.3.2 Data Pre-processing

The daily rainfall observations were normalised by rescaling each input vectors before fitting into the SOFM model. The min-max normalisation technique as shown in Equation (2) was used as it could minimise the effect of any overweighted computations on the input instances during the training phase.

$$z_{ij} = \frac{x_{ij} - min\{x_i\}}{max\{x_i\} - min\{x_i\}} \quad , \; 0 \leq z_{ij} \leq 1 \tag{2}$$

2.3.3 Data Training

2.3.3.1 Initialisation Phase

Parameters were initialised at random for all $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, M$ at iteration $s = 0$. The initialisation of the parameters of the connection weights between each of the m input node and the M node in the Kohonen map, and the learning rate of the SOFM model is defined in Equation (3) and Equation (4) respectively.

$$w_{ij}(0) \sim \text{Uniform}(0,1) \tag{3}$$

$$\eta(0) \sim \text{Uniform}(0,1) \tag{4}$$

2.3.3.2 Competition Phase:

For each of the training iteration s, all connected neurons were competed based on Equation (5) to obtain the best matching unit (BMU), θ based on the minimised Euclidian metrics.

$$D_j{}^{(\theta)} = \underset{1 \le j \le M}{\text{argmin}} \left\| z_i(s) - w_{ij}(s) \right\| \text{ where } D_j = \sqrt{\sum_{i=1}^{m} \left(z_i(s) - w_{ij}(s) \right)^2} \tag{5}$$

2.3.3.3 Cooperative Update and Learning Phase:

Based on the BMU that was derived from the competition phase, the radius of the neighbourhood region, N_θ was computed for each iteration as shown in Equation (6). The Gaussian neighbourhood function, $H_\theta(s)$, for the size of N_θ, around the BMU is defined in Equation (7), where T is the maximum iteration number, and θ^* is the closest node towards the BMU.

$$\delta(s) = \delta(0) e^{-\frac{s}{T}} \tag{6}$$

$$H_\theta(s) = e^{-\frac{1}{2}\left(\frac{\|\theta - \theta^*\|}{\delta(s)}\right)^2} \tag{7}$$

Equation (8) indicates the linear learning rate function, $\eta(s)$. Only the weights of the winning nodes (i.e. BMU) were updated.

$$\eta(s) = \frac{\eta(0)}{s} \tag{8}$$

The following pseudocode describes the updating process of the nodes in the Kohonen map. The corresponding neighbourhood nodes were activated whereas the nodes which has not been updated were deactivated.

if $j \in H_\theta(s)$,
then $w_{ij}(s+1) = w_{ij}(s) + \eta(s) H_\theta(s) \left[z_i(s) - w_{ij}(s) \right]$,
else $w_{ij}(s+1) = w_{ij}(s)$

For the case where input vectors consisting missing daily rainfall values were identified at each iteration s, the calculation in the competitive stage was ignored for the vectors. The corresponding weights will not be updated as shown in Equation (9).

$$w_{ij}(s+1) = w_{ij}(s) \tag{9}$$

The algorithm runs for each iteration s, until the training converges for all of the completely available daily rainfall data. The neighbourhood function contributed to the weight adjustments, where closer nodes towards the BMU were updated more frequently. After the weight vectors in the Kohonen map were fully updated and adjusted accordingly, a clear illustration for the rainfall distributions were provided in the form of feature maps. In the feature maps, similar clusters were identified by physically closely-arranged nodes. The

feature map was also constructed to help investigate the individual breakdown of rainfall intensities for each rainfall station.

2.3.4 Missing Data Estimation

The Gaussian neighbourhood function within the Kohonen map allowed a posteriori estimation due to the asymptotic convergence behaviour towards the arithmetic average value of the cluster around the BMU, $\overline{x}^{(\theta)}$. Consequently, the missing daily rainfall values were estimated utilising Equation (10).

$$\hat{x}_{MD} = \overline{x}^{(\theta)} \text{ in } N_\theta \tag{10}$$

2.4 Statistical Performance Metrics

The performance of the SOFM model in estimating the missing daily rainfall values was assessed based on the following statistical performance metrices, the root mean square error (RMSE) in Equation (11), and the coefficient of determination (R^2) in Equation (12). The calculations involved the N estimated daily rainfall values, $\hat{x}_k$ compared to the original observed daily rainfall values, x_k.

$$\text{Root Mean Square Error, } RMSE = \sqrt{\frac{\sum_{k=1}^{N}(\hat{x}_k - x_k)^2}{N}} \tag{11}$$

$$\text{Coefficient of Determination, } R^2 = \frac{\sum_{k=1}^{N}(\hat{x}_k - \overline{x}_k)^2}{\sum_{k=1}^{N}(x_k - \overline{x}_k)^2} \tag{12}$$

3 Results and Discussions

Figure 2 – Figure 10 illustrated the individual feature maps for each of the respective rainfall stations, extracted from the trained Kohonen map in the SOFM model. The colours illustrated in the feature maps indicates the rainfall intensities that were estimated. The warmer the colour temperature that was filled by the node, the lower the daily rainfall amount. On the contrary, the cooler the node colour temperature, the higher the daily rainfall amount recorded.

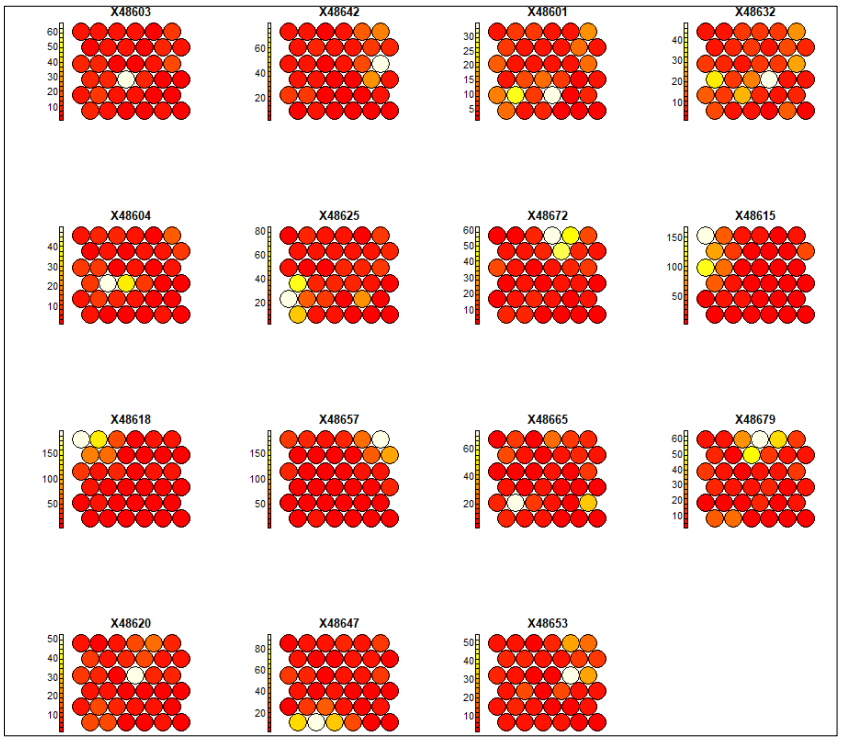

Fig. 2. Feature maps of the Kohonen map of the SOFM model for 10% missing data.

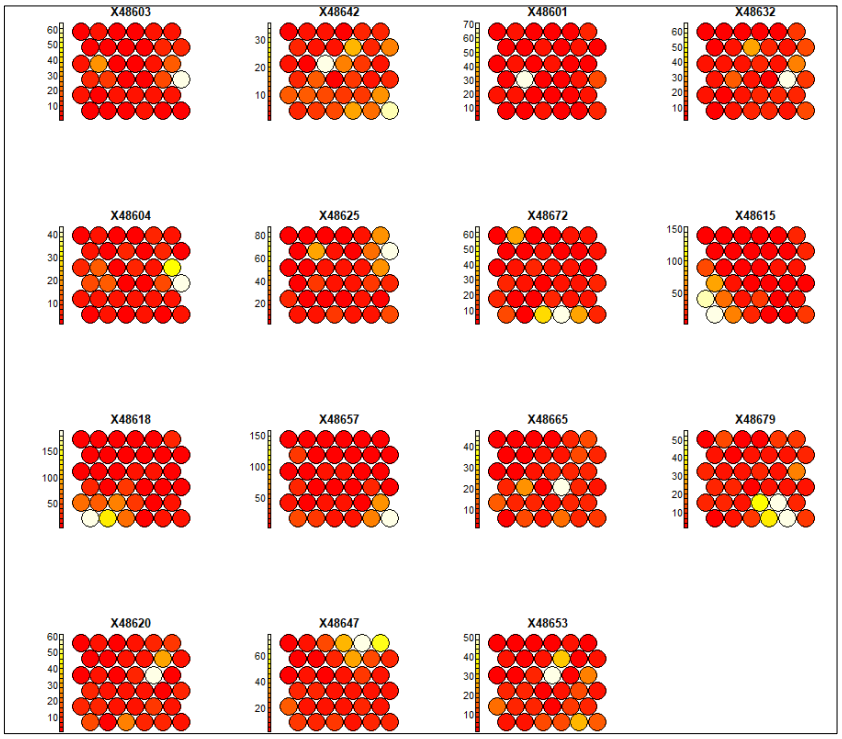

Fig. 3. Feature maps of the Kohonen map of the SOFM model for 15% missing data.

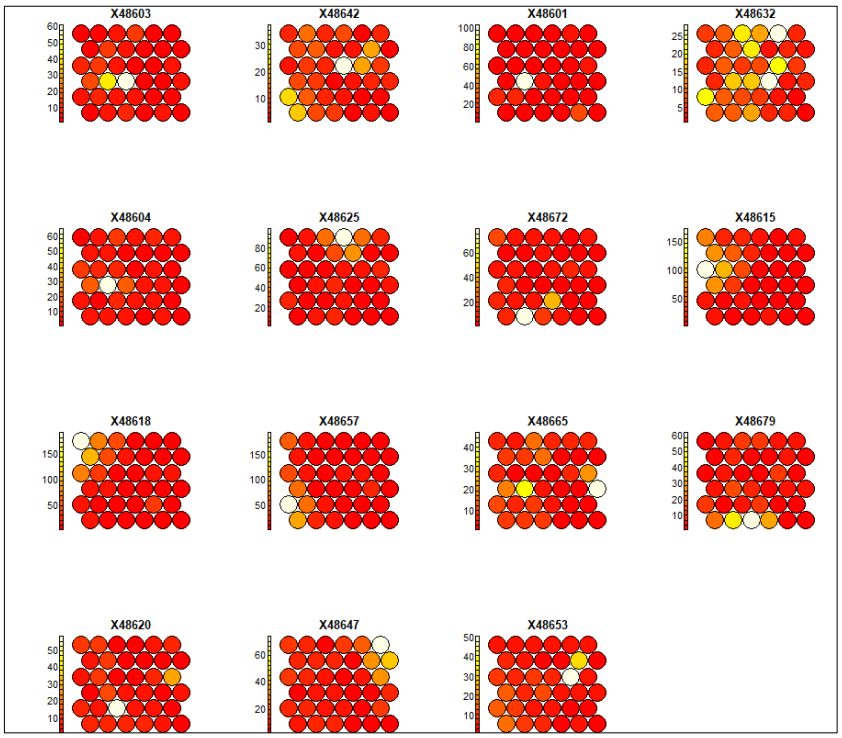

Fig. 4. Feature maps of the Kohonen map of the SOFM model for 20% missing data.

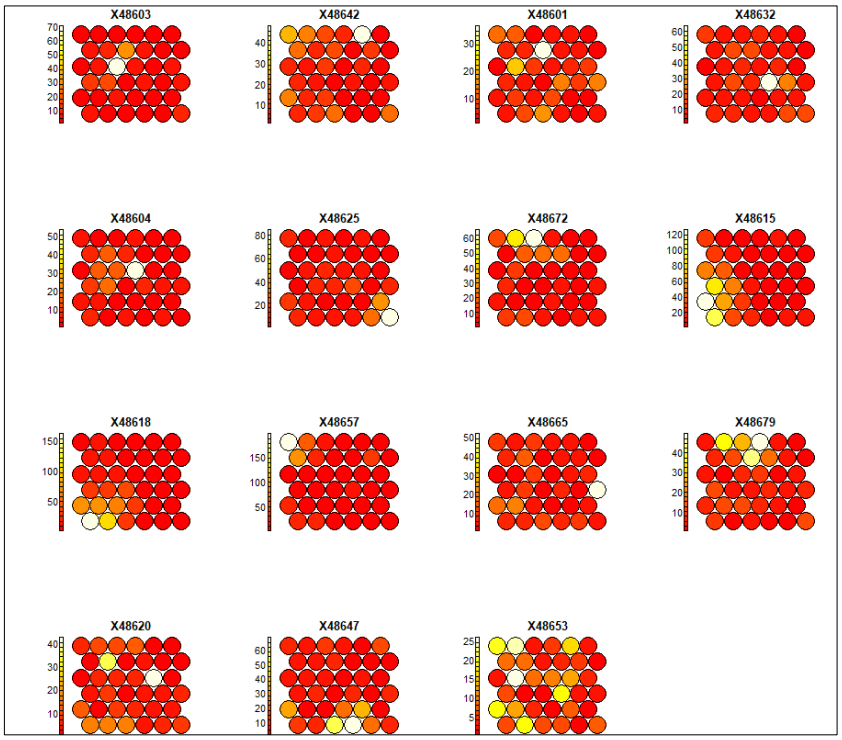

Fig. 5. Feature maps of the Kohonen map of the SOFM model for 25% missing data.

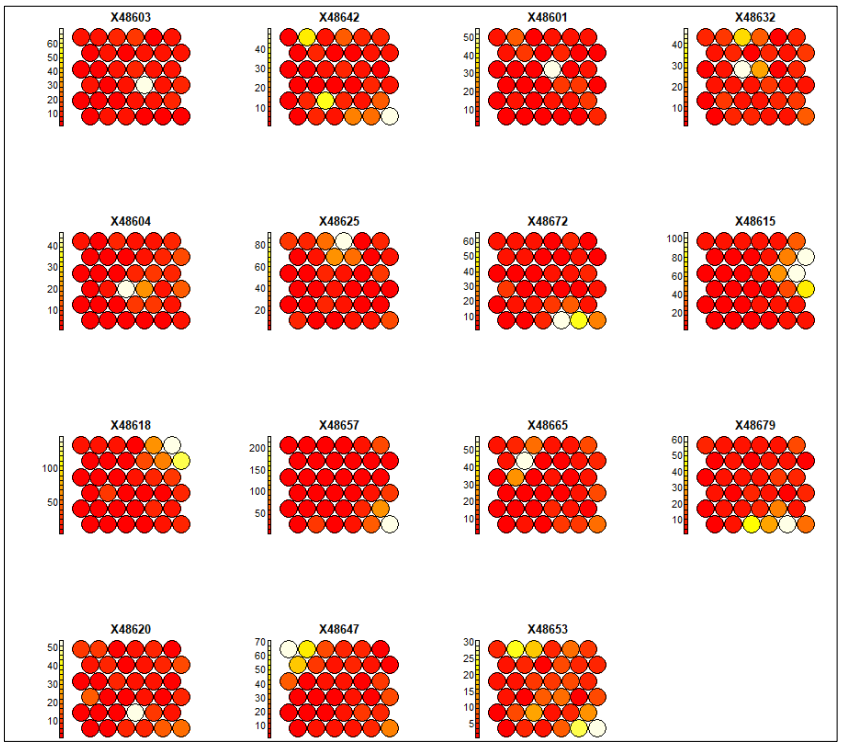

Fig. 6. Feature maps of the Kohonen map of the SOFM model for 30% missing data.

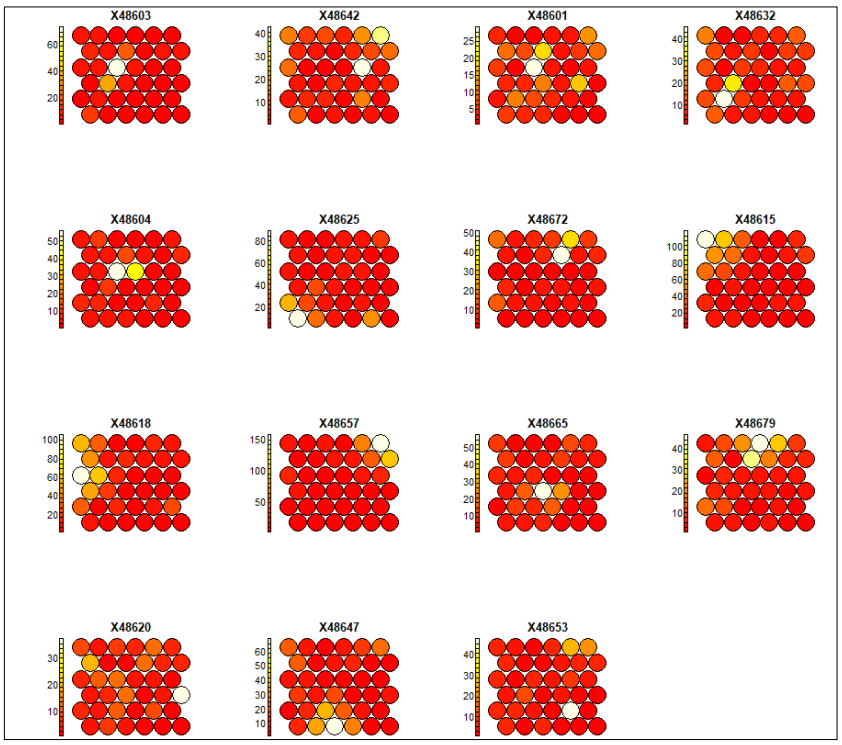

Fig. 7. Feature maps of the Kohonen map of the SOFM model for 35% missing data.

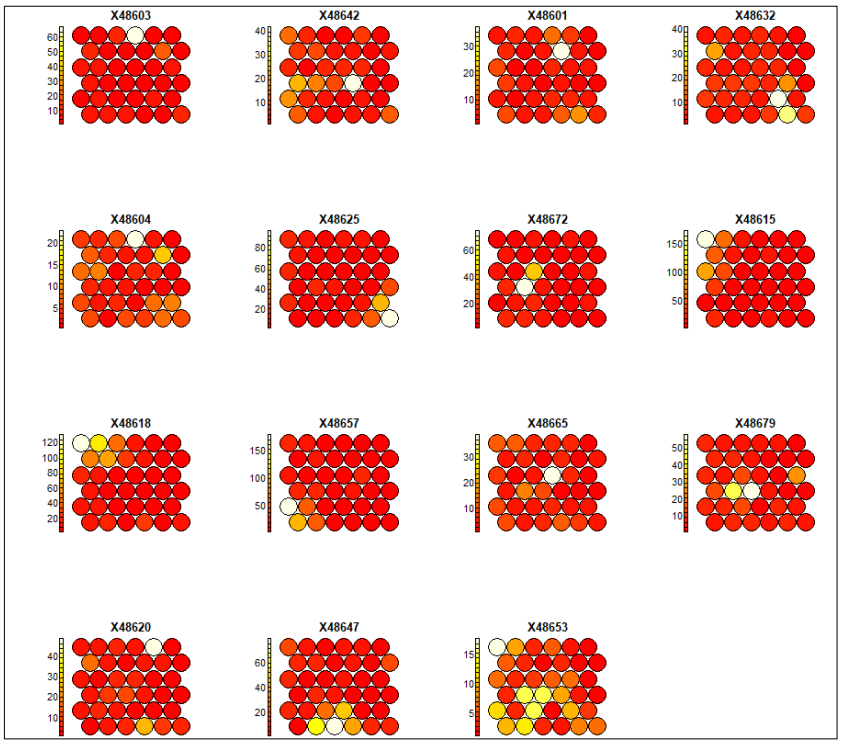

Fig. 8. Feature maps of the Kohonen map of the SOFM model for 40% missing data.

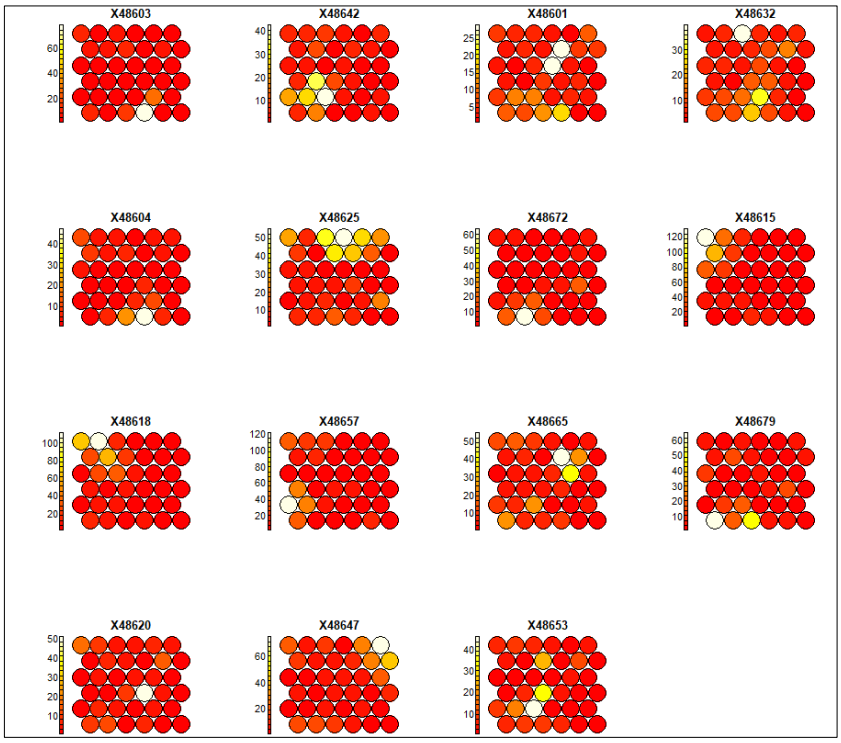

Fig. 9. Feature maps of the Kohonen map of the SOFM model for 45% missing data.

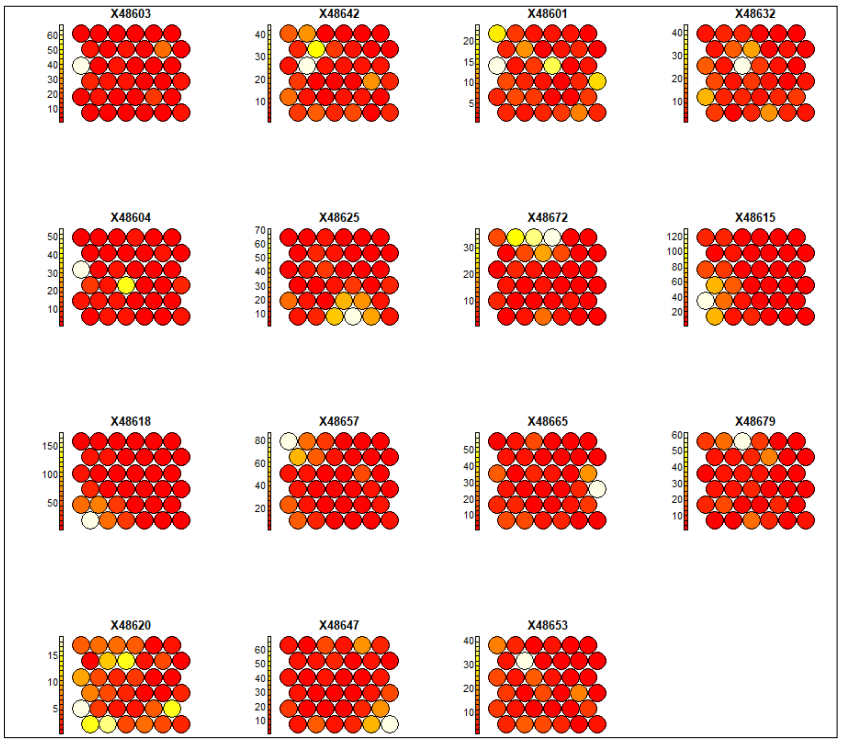

Fig. 10. Feature maps of the Kohonen map of the SOFM model for 50% missing data.

From the feature maps of the SOFM model, the largest range of rainfall amounts were recorded in the stations of Kuala Terengganu (X48618) and Kuantan (X48657). This reflected that these rainfall stations received a significantly large daily rainfall amount during the NEM period. The results could be justified by their locations which are both positioned at the near boundaries of the North East side of the Peninsular Malaysia. Besides, the rainfall stations are nearest to the South China Sea as shown in the Peninsular Malaysia geographical map. However, they are inhomogeneous due to the disparate projected patterns examined from the feature maps. On the other hand, the Senai (X48679) and Kluang (X48672) rainfall stations exhibited a somewhat similar rainfall pattern across all percentage of missing data although there was slight difference in the rainfall intensity. This could be attributed to the fact that the two stations were significantly closer to each other compared to the other rainfall stations. By careful inspection, the feature maps in Figure 4 have depicted that the Bayan Lepas (X48601) and Chuping (X48604) rainfall stations showed highly similar feature map projections when the missing rainfall data percentage is at 20%. Both of the feature maps in the lattice structure arrangement shared an identical node with the highest rainfall intensity (filled with white colour) which was located at the fourth row second column. Although they have similar patterns, the colour intensities beside the white nodes are slightly different. The feature map of the Chuping station has orange-coloured nodes beside the white node whereas the feature map of the Bayan Lepas station was surrounded by only red-coloured nodes. Besides, the maximum daily rainfall amount estimated under the Bayan Lepas rainfall station is up to 100 mm compared to the Chuping station with maximum observed value of 60 mm. The overall findings based on the feature maps visualisation of the SOFM Kohonen maps suggested that the 15 studied rainfall stations have distinct rainfall dynamics regardless of the missing daily rainfall data percentages. In other words, it could be concluded that the studied rainfall stations were inhomogeneous.

On top of the feature map visualisations, Table 2 provided the statistical performance metrics of the SOFM model across the different percentages of missing daily rainfall data.

Table 2. Statistical performance metrics of the SOFM for different percentage of missing data.

Missing Data Percentage (%)	RMSE (in mm)	R^2
10	15.9	0.915
15	17.5	0.865
20	18.8	0.731
25	19.2	0.680
30	20.0	0.617
35	20.8	0.546
40	22.2	0.522
45	22.7	0.504
50	29.3	0.499

Based on the statistical performance metrics, the SOFM model has the optimal performance for the smallest missing percentage of 10%. The outstanding estimation performance of the SOFM was reflected by the lowest RMSE of 15.9 mm and the highly significant R^2 value of 0.915. It was also evident that the highest missing data percentage of 50% yielded the poorest estimation performance. This was shown by the largest RMSE of 29.3 mm, and the lowest R^2 value of 0.499. Overall, the increased percentages of missing daily rainfall data led to an increase in the RMSE, and a decrease in the R^2 value. It is worth to note that although the SOFM had the worst performance when the missing data is 50%, it was able to secure the R^2 of approximately 50%. In particular, the significant R^2 value suggested that a significant proportion of the variability in the estimated missing daily rainfall data can be explained by the SOFM model. This is because a higher R^2 value reflects that the SOFM model provided a good fit to the missing data points when compared with the complete rainfall observations. In short, the result findings confirmed the outstanding estimation performance of the SOFM as it was capable of producing reliable estimates for the daily rainfall missing data despite of a large loss of available data (up to 50%).

4 Conclusion

The SOFM model provided visualisation on the estimated missing daily rainfall data where the conventional machine learning models were unable to produce due to the black-box property. Moreover, the unsupervised learning mechanism of the SOFM model enabled an unbiased estimation of the missing daily rainfall data without pre-assuming any relationships between the rainfall stations and the missing as well as observed daily rainfall data itself. In addition, the use of inhomogeneous rainfall stations in estimating the missing daily rainfall data ease researchers in the process of rainfall stations selections. This is because the successful missing data estimation performance exhibited by the SOFM model suggested that the SOFM worked well regardless of the rainfall station homogeneity. This is an added advantage since there are no pre-assumptions in selecting the rainfall stations.

From the perspective of the computed statistical performance metrics, the SOFM was shown to produce remarkable estimation results on the missing daily rainfall data, by achieving a significantly high R^2 value of approximately 92%, associated with a low error indicated by the lowest RMSE value of 15.9 mm. Across all different percentages of missing data, the SOFM model was also capable of producing reliable estimates for the missing daily rainfall data even though there were large numbers of data being missing. In conclusion, the SOFM model is indeed a reliable machine learning model that is highly useful and accurate in the missing daily rainfall data estimations. Additionally, the innovative feature maps generated by the SOFM model provided useful insights in terms of the rainfall stations pattern visualisation. It is recommended that for the future research directions, the study duration can be expanded to more rainfall stations, and also with the time period of more than 10 years in order to obtain a more precise data estimates. Furthermore, a diverse range of machine learning based metaheuristic optimisation algorithms such as the particle swarm optimisation (PSO), genetic algorithm (GA), the ant-lion optimisation (ALO), etc., which could provide optimisation on the model performance.

This research was funded by the Universiti Tunku Abdul Rahman Research Fund (IPSR/RMC/UTARRF/2023-C1/L01). The authors would like to thank the Lee Kong Chian, Faculty of Engineering and Science, Universiti Tunku Abdul Rahman for all the technical support provided for this research.

References

1. W.S. Loh, R.J. Chin, L. Ling, S.H. Lai, E.Z.X. Soo, Mathematics, **9**, 3141 (2021)
2. L. Ling, Y. Zulkifli, J.L. Ling, Mathematics, **9**, 812 (2021)
3. World Bank and Asian Development Bank's 2021 publication Climate Risk Country Profile – Malaysia and the UN's 2021 publication Disaster Risk Reduction in Malaysia: 2020 Status Report.
4. A.A.G. Nadiatul Adilah, H. Hannani, *Comparison of methods to estimate missing rainfall data for short term period at UMP gambang2021*, in IOP Conf. Ser.: Earth Environ. Sci. **682**, 012027 (2021)
5. W. Angkool, M. Waqas, P. Dechpichai, P.T. Hlaing, S. Ahmad, U.W. Humphries, Imputation of missing daily rainfall data; A comparison between artificial intelligence and statistical techniques, MethodsX. **11**, 102459 (2023)
6. R.S.V. Teegavarapu, Hydrol. Sci. J., **59**(11), 2009 – 2026 (2014)
7. C.W. Dawson, W. Robert. Prog. Phys. Geogr., **25**(1), 80 – 108 (2001)
8. R.J. Chin, S.H. Lai, W.S. Loh, L. Ling, E.Z.X. Soo, *Assessment of Inverse Distance Weighting and Local Polynomial Interpolation for Annual Rainfall: A Case Study in Peninsular Malaysia.* Eng. Proc. **38**(1), 61 (2023)
9. P. Wolski, C. Jack, M. Tadross, L. van Aardenne, C. Lennard, Clim. Dyn., **50**, 479 – 492 (2018)
10. H. Moradkhani, K. Hsu, H.V. Gupta, S. Sorooshian, J. Hydrol., **295**(1 – 4), 246 – 262 (2004)
11. M.K. Ho, Y. Zulkifli, Int. J. Comput. Appl., **48**(5) (2012)
12. R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, 2nd edn. Hoboken, NJ. (John Wiley & Sons, Inc., 2002)
13. P. Sharma, U. Mutreja. Int. J. Soft Comput. Eng., **2**, 276 – 278 (2013)