Imputation of Rainfall Data Using the Sine Cosine Function Fitting Neural Network

Po Chan Chiu[1,2*], Ali Selamat[1,3,4*], Ondrej Krejcar[4], King Kuok Kuok[5], Enrique Herrera-Viedma[6], Giuseppe Fenza[7]

[1] School of Computing, Faculty of Engineering & MagicX (Media and Games Center of Excellence), Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor (Malaysia)
[2] Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak (Malaysia)
[3] Malaysia Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia Kuala Lumpur, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur (Malaysia)
[4] Faculty of Informatics and Management, University of Hradec Kralove, Rokitanského 62, 500 03 Hradec Kralove (Czech Republic)
[5] Faculty of Engineering, Computing and Science, Swinburne University of Technology Sarawak Campus, 93350 Kuching, Sarawak (Malaysia)
[6] Andalusian Research Institute Data Science and Computational Intelligence, University of Granada, 18071 Granada (Spain)
[7] Dipartimento di Scienze Aziendali-Management & Innovation Systems (DISA-MIS), University of Salerno, 84084 Fisciano (Italy)

Received 13 September 2020 | Accepted 2 July 2021 | Published 12 August 2021

unir
LA UNIVERSIDAD EN INTERNET

Abstract

Missing rainfall data have reduced the quality of hydrological data analysis because they are the essential input for hydrological modeling. Much research has focused on rainfall data imputation. However, the compatibility of precipitation (rainfall) and non-precipitation (meteorology) as input data has received less attention. First, we propose a novel pre-processing mechanism for non-precipitation data by using principal component analysis (PCA). Before the imputation, PCA is used to extract the most relevant features from the meteorological data. The final output of the PCA is combined with the rainfall data from the nearest neighbor gauging stations and then used as the input to the neural network for missing data imputation. Second, a sine cosine algorithm is presented to optimize neural network for infilling the missing rainfall data. The proposed sine cosine function fitting neural network (SC-FITNET) was compared with the sine cosine feedforward neural network (SC-FFNN), feedforward neural network (FFNN) and long short-term memory (LSTM) approaches. The results showed that the proposed SC-FITNET outperformed LSTM, SC-FFNN and FFNN imputation in terms of mean absolute error (MAE), root mean square error (RMSE) and correlation coefficient (R), with an average accuracy of 90.9%. This study revealed that as the percentage of missingness increased, the precision of the four imputation methods reduced. In addition, this study also revealed that PCA has potential in pre-processing meteorological data into an understandable format for the missing data imputation.

Keywords

Imputation, Missing Rainfall Data, Principal Component Analysis (PCA), Sine Cosine Neural Network, Deep Learning.

DOI: 10.9781/ijimai.2021.08.013

I. Introduction

RAINFALL is a critical component of the hydrological cycle. Numerous hydrological research areas, such as flood forecasting [1], flood risk assessment [2], rainfall forecasting [3], climate variability analysis [4], and water resources modeling [5], require reliable and complete rainfall data series. However, hydrological data analysis is challenging due to the presence of missing rainfall data. For this reason, data imputation has attracted a great deal of attention from researchers to fill in the missing values with approximations. The traditional imputation approaches include listwise deletion [6], arithmetic mean and median imputation [7], and multiple imputations [8]. However, these methods are time-consuming and less accurate [9].

In recent years, numerous artificial neural network (ANN) studies have used historical rainfall data series from nearest neighbor stations to treat the problems of missing data [10]-[12]. More efficient algorithms, such as the Levenberg-Marquardt backpropagation algorithm [13], the Gaussian mixture model-based K-nearest neighbor (GMM-KNN) algorithm [14], and the Bayesian principal component analysis (BPCA) [15] have been applied to impute the missing values in water resource engineering.

* Corresponding author.
E-mail addresses: pcchiu@unimas.my (P. C. Chiu), aselamat@utm.my (A.Selamat).

Although ANNs have been applied to treat the problem of missing data, ANNs tend to be trapped in local optima as it smoothly converges towards local minima rather than global minima. To overcome this, several novel approaches have been combined with ANNs to improve the performance of the estimation results. The sine cosine algorithm (SCA) is a metaheuristic technique developed by Mirjalili [16] to solve optimization problems using the sine and cosine trigonometric functions. SCA has been successfully applied in modal dimensional [17], short-term hydrothermal scheduling [18], support vector regression [19], and the traveling salesman problem [20]. To the best of the authors' knowledge, there is no existing sine cosine neural network that focuses on missing rainfall data imputation.

Furthermore, the use of raw hourly rainfall data from nearest neighbor stations could be unreliable for the prediction of the missing data of the target station. The long dry periods contain long sequences of zero values at the beginning, middle, or end of the records, in which rain does not usually fall every hour. Modeling long dry rainfall periods poses challenges such as underestimation or overestimation of the length of long dry periods [21] and [22]. As a result, a neural network is not able to estimate the missing rainfall value based on hourly rainfall datasets accurately. Hence, the hourly rainfall dataset needs to be combined with other non-precipitation data for the estimation of missing rainfall data.

According to Kashiwao et al. [23], rainfall is caused by a variety of meteorological conditions, and the mathematical model for it is non-linear. The meteorological data have different units of measurement and accuracy. Thus, the meteorological data need to be pre-processed prior to imputation. Normalization is the most commonly used approach. Yen [24] applied a *mapminmax* approach to normalizing the meteorological parameters in the study, while Chhetri et al. [25] normalized the weather parameters using a min-max scaler. In addition, Grange [26] proposed using a random forest machine learning algorithm for meteorological normalization to detect interventions in an air quality time series. According to Kashiwao et al. [23], the investigation into the method used to choose meteorological data is needed because suitable data can vary among prediction points due to the difference in the effect of conditions, such as altitude, ocean current, and airflow. For this reason, this paper proposes using principal component analysis (PCA) as a novel pre-processing mechanism to extract the core relationships in the meteorological data. PCA is used to identify patterns in data and express the similarities and differences of the data [27]. PCA has been used in many studies to isolate independent factors (principal components) that significantly explain the variation of a dependent variable [28]-[32]. However, the compatibility of both non-precipitation and precipitation as input has been given less attention in previous studies. Therefore, we propose using PCA as a novel pre-processing tool for meteorological data and introduce the combination of significant principal components (PCs) and rainfall data from nearest neighbor gauging stations as the input for the estimation of missing rainfall values.

The contributions of this paper are the following:

- To introduce a pre-processing mechanism for non-precipitation data by using principal component analysis (PCA).
- To propose a sine cosine function fitting neural network (SC-FITNET) imputation that focuses on missing time series data.
- To evaluate the performances of sine cosine function fitting neural network (SC-FITNET) imputation with the state-of-art models for infilling missing rainfall values at different percentages of missingness.

II. Methodology

The proposed methodology employed in this study consists of two main phases, as shown in Fig. 1. The phases are the data preparation phase and the missing data imputation phase.

A. Phase 1: Data Preparation

In this study, the data preparation phase attempts to transform raw data into an understandable format prior to the missing data imputation. The data preparation phase involves data pre-processing and data integration. Due to the variety of measurement units, the raw meteorological data must be pre-processed. For example, the values of mean surface wind (direction) are stored at 00°, 010°, ..., 058°. These characters are considered noise in the data because the neural network could not understand and interpret those characters accurately.

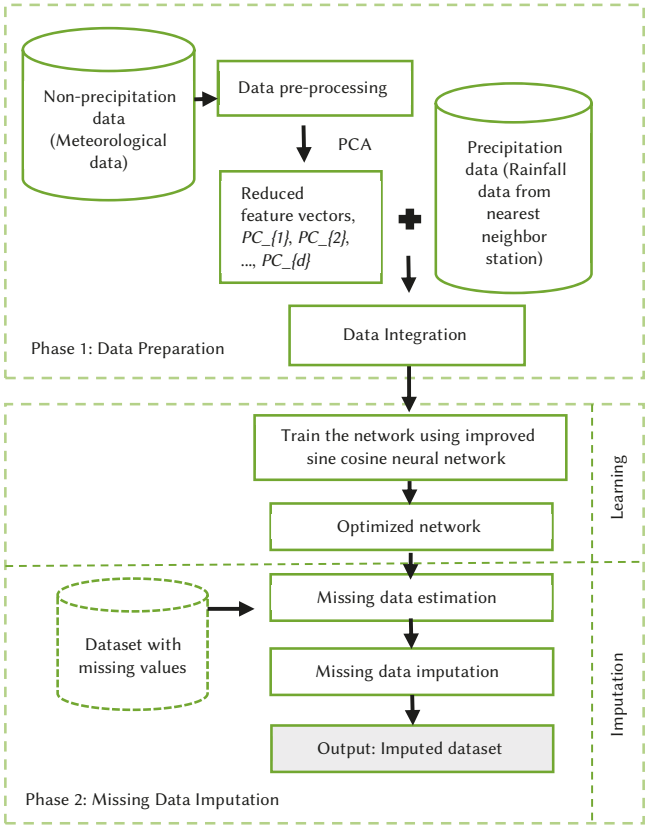

Fig. 1. The proposed methodology of missing data imputation.

In the related literature, the advantages of PCA are able to reveal hidden structure in the dataset, detect outliners, and filter out the noise in data [33]. In addition, PCA is one of the most used approaches to pre-process the weather [28] and meteorological data [30]. Therefore, PCA was used to pre-process the raw meteorological data.

PCA was proposed by Pearson [34] and formalized by Hotelling [35]. Using PCA, these meteorological data were transformed into a smaller number of variables. PCA reduces the number of meteorological features by constructing a new and smaller number of variables that capture a significant portion of the original meteorological features. The pre-process of meteorology data starts with normalizing the variables by subtracting the mean from each data point. Next, the covariance and correlation between every pair of variables (meteorological features) were calculated based on the following equations [27] and [36]:

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} \quad (1)$$

where, $cov(x, y)$ is the covariance of the variables x and y, x_i and y_i are the independent variable of observations, $\bar{x}$ and $\bar{y}$ are the mean values of the variables x_i and y_i, respectively and n is the number of data points in the observations.

$$r(x,y) = \frac{cov(x,y)}{s_x s_y} \tag{2}$$

where, $r(x, y)$ is the correlation of the variables x and y, s_x is the sample standard deviation of the random variable x, and s_y is the number of data points in the observations.

Then eigenvector and eigenvalue of the matrix are obtained as follows:

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n,1} & A_{n,2} & \cdots & A_{n,n} \end{bmatrix} \tag{3}$$

$$Av = \lambda v \tag{4}$$

The eigenvector v of each variable can be obtained by identifying the determinant of its characteristic polynomial as follows:

$$(A - \lambda I)v = 0 \tag{5}$$

The eigenvalue can be formulated using the following Equation:

$$p(\lambda) = |A - \lambda I| \tag{6}$$

After these steps, the principal components (*PC_{1}, PC_{2}, …, PC_{d}*) can be determined. The first principal component accounts for the highest variance in the meteorological dataset, followed by the second principal component for the next highest variance. This continues until the total of the principal components is equal to the number of features in the meteorological dataset.

The last step is to compute the feature vector. A matrix M of dimensions $n \times d$ is represented as

$$M = \begin{bmatrix} f_{1,1} & f_{1,2} & f_{1,3} & \cdots & f_{1,d} \\ f_{2,1} & f_{2,2} & f_{2,3} & \cdots & f_{2,d} \\ f_{3,1} & f_{3,2} & f_{3,3} & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{n,1} & f_{n,2} & f_{n,3} & \cdots & f_{n,d} \end{bmatrix} \tag{7}$$

where, f_{ij} is a reduced feature vector from $n \times n$ original data to size $n \times d$, n is the number of data points in the observations, and d is the number of principal components.

The final output of the PCA is combined with the raw rainfall data from the nearest neighbor gauging stations and then used as the input to the neural network for missing data imputation.

B. Phase 2: Missing Data Imputation

The missing data imputation phase consists of two sub-phases, namely learning and imputation. In the learning sub-phase, the combined dataset from phase 1 will be used as an input to the neural network training. By using the ANN approach, the neural network is trained and optimized to learn the complex and non-linear relationships between the features in the dataset. The output of the learning sub-phase is an optimized network with a set of optimal network weights and biases. Next, the imputation sub-phase involves missing data estimation using the optimized network. During the missing data imputation, the estimated missing data are imputed into the missing values in the dataset. Hence, the final output of this phase is the imputed database.

III. Imputation Methods

Artificial neural networks (ANNs) based rainfall and runoff (R-R) modeling were first applied in the early 1990s. ANNs learn complex and non-linear relationships that are difficult to model using statistical approaches. Hence, in this study, four ANN models are employed to estimate the missing time series values.

A. Feedforward Neural Network (FFNN)

The feedforward neural network (FFNN) model is the simplest type of ANNs [37]. The architecture of the FFNN network consists of p -many inputs (input neurons), a single hidden layer with q-many hidden neurons, and a single output. A simulation for estimation of the missing rainfall data using FFNN was carried out with ten neurons in the hidden layer. The activation functions for the hidden layer and output layer are tan-sigmoid and purelin, respectively.

B. Sine Cosine Function Fitting Neural Network (SC-FITNET)

The function-fitting neural network (FITNET) is a feedforward network that forms a generalization of the input and output relationship. FITNET produces an associated set of target outputs, with tan-sigmoid transfer function in the hidden layers and linear transfer function in the output layer. The FITNET model was trained with two hidden layers; a first hidden layer with 15 neurons and a second layer with three neurons.

To improve the performance of missing data prediction, the FITNET model is optimized by the sine cosine algorithm (SCA). The improved neural network is therefore named as sine cosine function fitting neural network, abbreviated as SC-FITNET. The sine cosine algorithm (SCA) is a metaheuristic optimization technique introduced by Mirjalili [16] to solve continuous optimization problems. One of the most significant advantages of SCA is its simplicity, as reported by Qu et al. [17]. SCA has fewer parameters that need to be fine-tuned compared to other algorithms. The capability of SCA in missing rainfall data imputation has not yet been explored. Hence, the SCA is employed to train the FITNET model for missing data prediction.

First, the network is trained using a function-fitting neural network to identify and learn the relationships between features in the dataset. Then, the SCA is employed to optimize the search solutions by determining the optimal network weights and biases.

SCA starts the optimization process with a set of search solutions, X. The set of search solutions is initialized randomly and repeatedly evaluated by an objective function. The objective of the training is to minimize the prediction error. The evaluation of the training is measured by the mean square error (MSE) as follows [38]:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y - \tilde{y})^2 \tag{8}$$

where N is the number of observations, y is the actual value, and $\tilde{y}$ is the predicted value.

Next, the search solution is improved by the position-updating function in Equation (9)[16]. The SCA updates the best solutions obtained and denotes it as a destination point, P.

$$X_i^{t+1} = \begin{cases} X_i^t + r1 * \sin(r_2) * \left| r_3 P_i^t - X_{i,}^t \right|, & r_4 < 0.5 \\ X_i^t + r1 * \cos(r_2) * \left| r_3 P_i^t - X_{i,}^t \right|, & r_4 \geq 0.5 \end{cases} \tag{9}$$

where, X_i is the position vector of the current solution in the i^{th} dimension, t is the current iteration, P_i is the destination solution, r_1, r_2, r_3, r_4 are random variables, and the r_4 value is between 0 and 1.

As seen in (9), there are four parameters in SCA, namely r_1, r_2, r_3 and r_4. The parameter r_1 is the movement direction parameter that determines the region of the next solution, which is updated using (10). The parameter r_2 identifies the movement of forwards or outwards P_i

within the value of 0 and 2π. Next, the parameter r_3 is the random weights of P_i with a value either less than 1 or greater than 1. The parameter r_4 is used to switch between the sine and cosine functions.

$$r_1(t) = a * \left(1 - \frac{t}{t_{max}}\right) \tag{10}$$

where, t is the current iteration, t_{max} is the maximum iteration of SCA, and a is a constant.

As the iteration of SCA increases, the ranges of sine and cosine in the position-updating functions are updated to optimize the local search, as shown in Line 7 of Algorithm 1. Then, the best network weights and biases are updated to improve the network model. The execution of the search solution will be halted if the network has achieved the minimum error or reached the maximum network epochs. Next, given the optimized network, the network model is tested with another dataset of the same format to predict the missing rainfall data. Then, the estimated missing rainfall data are imputed into the missing dataset. The proposed SC-FITNET imputation is presented in Algorithm 1.

Algorithm 1: The proposed sine cosine function fitting neural network (SC-FITNET) imputation

Input: Pre-processed meteorology and nearest neighbor rainfall

1. **Do**
2. Select random search agents (solutions) (X) and SCA parameters (r_1, r_2, r_3 and r_4)
3. **Do**
4. Evaluate each of the search agents by the objective function
5. Update the best solution obtained so far (P)
6. Update the parameters r_1, r_2, r_3 and r_4
7. Update the position of search agents using Equation (9)
8. **While** (t < maximum number of iterations)
9. **Return** the best solution (P) obtained as the global optimum solution
10. Track the best network
11. Update training state
12. **While** (MSE > the minimum error) or (epoch < maximum number of epochs)
13. Use the optimized network
14. Train the optimized net for another dataset of the same format
15. **Do**
16. Impute the estimated values into the missing value
17. **While** (there is missing value)

Output: Imputed rainfall dataset

Note: The algorithm in the dotted line box was adapted from Mirjalili [16]

In addition, different values of the parameters are introduced to the SC-FITNET. The parameters are tuned based on the try and error method. The parameter settings are outlined in Table I.

TABLE I. The SC-FITNET Parameters

Parameters for SC-FITNET	Value
a	2
Search agents	30
Max number of epochs	1000
Max iteration of SCA	500

C. Sine Cosine Feedforward Neural Network (SC-FFNN)

The third model evaluated was a sine cosine feedforward neural network (SC-FFNN). The adaptation of the sine cosine algorithm into the feedforward neural network is employed to improve the accuracy of missing rainfall data imputation. The model was trained with ten neurons in the hidden layer. The SC-FFNN applied the same parameters setting, as in Table I.

D. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a recurrent deep neural network model [39]. Recent studies have successfully applied LSTM based deep learning models for time series forecasting [40], data augmentation [41] and sequence labeling [42]. Hence, we developed a LSTM multivariate time series model to predict the missing values. The LSTM model consists of five layers; an input layer, two layers of LSTM, a fully connected dense layer, and an output layer, as illustrated in Fig. 2. The two LSTM layers are employed to model the time series relationship, while the fully connected layer takes the output of the LSTM layers to a final missing data prediction.

After data pre-processing, the data are reshaped into a multivariate format for the LSTM models. The activation function used in this model was the default tanh, Adam optimizer, 20 epochs of training with a batch size of 32, GPU execution environment, two hidden layers of 120 neurons each and one-time delay handling the prediction of missing time series.

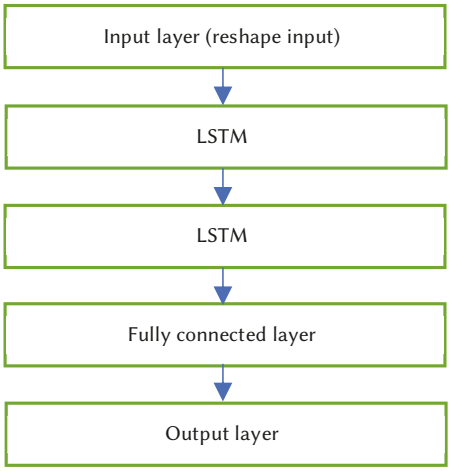

Fig. 2. The architecture of a LSTM network for multivariate time series prediction.

IV. Materials and Methods

A. Study Area

The selected study area for this study is Sungai Merang, or the Merang River gauge station, approximately 80 km from Kuching City, Sarawak, Malaysia. Sungai Merang is one of the five rainfall gauge stations in the Bedup River catchment, as shown in Fig. 3. Its nearest neighbor gauge stations over the basin are Bukit Matuh (BM), Semuja Nonok (SN), Sungai Busit (SB) and Sungai Teb (ST). The surface areas of the five rainfall gauge stations are SM: 8.550 km^2; BM: 8.075 km^2; SN: 7.600 km^2, SB: 8.075 km^2 and ST: 15.320 km^2.

The primary vegetation in this area is paddy and fruit plantation. The area is mostly covered with clayey soils and partly covered with coarse loamy soil. The soil texture enhances the infiltration rate but reduces the surface runoff. Hence, the water supply plan for paddy irrigation is crucial and extremely important for the village. However, the water supply plan and hydrological data analysis are challenging due to the presence of missing rainfall values at the Sungai Merang gauge station. Therefore, this study focuses on the missing rainfall data imputation at that gauge station.

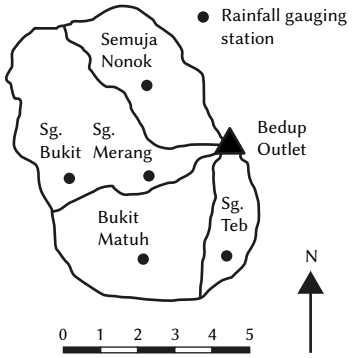

Fig. 3. Sungai Merang and its nearest neighbor gauging stations [3].

B. Meteorological Data

The meteorological data for Kuching station was acquired from the Malaysian Meteorological Department, as shown in Table II [43]. In this study, ten types of meteorological data were collected: date, time, the pressure at mean sea-level (MSL), dry-bulb temperature, relative humidity, mean surface wind (direction), mean surface wind (speed), rainfall duration, rainfall amount and cloud cover.

TABLE II. Meteorological Data From Kuching Station, Sarawak

Meteorological data	Measurement Unit
Date	YYMMDD
Time	MST
Pressure MSL	Hpa
Dry Bulb Temperature	° C
Relative Humidity	%
Mean Surface Wind (direction)	°
Mean Surface Wind (speed)	m/s
Rainfall Duration	min
Rainfall Amount	mm
Cloud Cover (cloud amount)	Oktas

C. Rainfall From Nearest Neighbor Stations

The rainfall data from the Sungai Merang gauging station and its nearest neighbor gauging stations were collected from the Department of Irrigation and Drainage, Sarawak, as shown in Table III [44]. Overall, the correlation coefficients between the Sungai Merang station and each of the neighbor stations are greater than 0.8 and located within a radius range of 5 km. Since the Sungai Merang gauging station exhibits a high correlation coefficient with its nearest neighbor stations, the complete rainfall data series from the four neighbor stations of the corresponding hour, day, month and year are used to predict the missing values of Sungai Merang's rainfall data. Based on the availability of continuous and complete data (without missing values) for the five gauging stations, this study analyzed the observed hourly rainfall data from the year 2002 until 2003. With a sample size of 11,680 complete records, the neural networks were trained with a training length of 8180 and tested with datasets of 3500 records. In [45]-[48], the data were randomly deleted and removed from the testing datasets. Hence, for the preparation of missing values in rainfall data, this study employed a rate-based approach [49] in which 10%, 20%, 30%, 40%, and 50% were randomly removed from the testing datasets. In total, two sets of testing data were prepared for each percentage of the missingness. In this study, the missing data were categorized as missing completely at random (MCAR) [50] because the presence of missing rainfall data at the Sungai Merang gauge station is not affected by the data in that area or any nearby area.

TABLE III. The Sungai Merang Gauging Station and Its Nearest Neighbor Gauging Stations

Station Name	Latitude	Longitude	Distance from Sg Merang (km)	Correlation Coefficient
Sungai Merang	001 05 40	110 36 25	-	-
Bukit Matuh	001 03 50	110 35 35	3.88	0.8558
Semuja Nonok	001 06 25	110 35 50	2.10	0.8647
Sungai Busit	001 05 25	110 34 40	3.44	0.8676
Sungai Teb	001 03 15	110 37 00	4.37	0.8046

D. Data Input Description

The number of data inputs, p, to the missing data imputation model was based on the number of cumulative principal components (*PC_{1}*, *PC_{2}*, …, *PC_{d}*) and raw rainfall data from the nearest neighbor stations.

$$input\ p\{d\} = cummulative\ of\ PC\_\{d\}, NNS1, NNS2, NNS3, NNS4 \quad (11)$$

where, PC is the principal component (s), $\{d\}$ is the number of principal components, and $NNS1$, $NNS2$, $NNS3$, $NNS4$ are the complete rainfall from the four nearest neighbor stations (NNS).

E. Performance Measures

The performances of the two imputation methods are measured by the mean absolute error (MAE), root mean square error (RMSE), and correlation coefficient (R).

- Mean absolute error (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|O_i - T_i| \quad (12)$$

- Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(O_i - T_i)^2}{N}} \quad (13)$$

- The correlation coefficient (R)

$$R = \frac{\sum_{i=1}^{N}(T_i - \bar{T})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{N}(T_i - \bar{T})^2 (O_i - \bar{O})^2}} \quad (14)$$

where N is the total number of observations, O_i is the actual values of observations, $\bar{O}$ is the mean values of the actual observations, T_i is the imputed values, and $\bar{T}$ is the mean of the imputed values.

V. Experiment

The proposed SC-FITNET missing data imputation was compared with the FFNN imputation, the SC-FFNN imputation and LSTM multivariate time series imputation using a combination input p of the meteorological data series (cumulative PC) and rainfall data series from nearest neighbor stations. A different number of inputs p was introduced, from $p1$ to $p10$, to determine the significant input p to the neural network. The average result gave the minimum MAE and RMSE measures, but the highest measure of R was chosen as the significant input p. For better evaluation of the proposed algorithm, we tested the imputation algorithms on two missing datasets. For each missing dataset, all the imputation algorithms were executed with 30 independent runs over each input p at different missing data rates (10%, 20%, 30%, 40%, and 50%). The average values of the performance measures for FFNN, SC-FFNN, SC-FITNET, and LSTM imputation, respectively, over two missing datasets, are presented in the following sub-sections.

TABLE IV. Comparison of MAE, RMSE, and R Values for FFNN Imputation at Different Percentages of Missingness

Input	MAE (mm)						RMSE (mm)						R					
P	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG
P1	0.113	0.217	0.275	0.398	0.497	0.300	1.029	1.454	1.338	1.741	1.976	1.508	0.920	0.834	0.863	0.789	0.711	0.823
P2	0.120	0.237	0.316	0.445	0.572	0.338	0.994	1.587	1.495	1.810	2.517	1.681	0.929	0.861	0.882	0.806	0.761	0.848
P3	0.134	0.271	0.362	0.514	0.647	0.386	1.043	1.629	1.515	2.017	2.350	1.711	0.925	0.855	0.872	0.803	0.754	0.842
P4	**0.108**	**0.212**	**0.280**	**0.397**	**0.497**	**0.299**	**0.871**	**1.210**	**1.103**	**1.435**	**1.638**	**1.251**	**0.947**	**0.896**	**0.914**	**0.856**	**0.808**	**0.884**
P5	0.148	0.286	0.387	0.533	0.686	0.408	1.186	1.757	1.678	1.887	2.427	1.787	0.919	0.865	0.884	0.818	0.774	0.852
P6	0.153	0.316	0.424	0.573	0.762	0.446	1.160	2.658	2.585	2.349	3.949	2.540	0.920	0.862	0.877	0.812	0.764	0.847
P7	0.150	0.298	0.406	0.571	0.708	0.426	1.003	1.424	1.365	1.724	1.920	1.487	0.927	0.855	0.871	0.808	0.759	0.844
P8	0.148	0.290	0.391	0.549	0.690	0.413	1.070	1.513	1.393	1.773	2.061	1.562	0.917	0.841	0.868	0.792	0.723	0.828
P9	0.164	0.345	0.452	0.610	0.786	0.472	1.219	2.891	2.783	2.285	3.547	2.545	0.897	0.795	0.820	0.752	0.680	0.789
P10	0.166	0.313	0.424	0.588	0.742	0.447	1.056	1.484	1.381	1.767	2.024	1.542	0.919	0.847	0.868	0.798	0.727	0.832

TABLE V. Comparison of MAE, RMSE, and R Values for SC-FFNN Imputation at Different Percentages of Missingness

Input	MAE (mm)						RMSE (mm)						R					
P	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG
P1	**0.072**	**0.135**	**0.163**	**0.244**	**0.301**	**0.183**	0.831	**1.108**	**0.952**	**1.344**	**1.473**	**1.142**	0.951	0.912	**0.936**	**0.874**	**0.842**	**0.903**
P2	0.087	0.165	0.214	0.310	0.388	0.233	**0.830**	1.108	1.014	1.375	1.534	1.172	**0.952**	**0.914**	0.930	0.870	0.837	0.901
P3	0.106	0.210	0.280	0.400	0.503	0.300	0.887	1.214	1.126	1.514	1.725	1.293	0.944	0.897	0.913	0.849	0.812	0.883
P4	0.116	0.228	0.296	0.427	0.529	0.319	0.957	1.305	1.140	1.546	1.739	1.337	0.933	0.875	0.905	0.832	0.785	0.866
P5	0.115	0.221	0.298	0.420	0.528	0.316	0.877	1.170	1.064	1.438	1.617	1.233	0.946	0.902	0.920	0.855	0.812	0.887
P6	0.113	0.220	0.296	0.420	0.522	0.314	0.869	1.158	1.038	1.425	1.591	1.216	0.947	0.904	0.925	0.857	0.815	0.889
P7	0.122	0.241	0.320	0.445	0.562	0.338	0.972	1.661	1.552	1.582	2.113	1.576	0.930	0.863	0.882	0.826	0.767	0.854
P8	0.142	0.279	0.381	0.537	0.665	0.401	1.019	1.412	1.324	1.733	1.920	1.481	0.929	0.876	0.897	0.829	0.786	0.863
P9	0.128	0.250	0.334	0.468	0.589	0.354	0.933	1.240	1.127	1.502	1.695	1.299	0.939	0.889	0.909	0.841	0.788	0.873
P10	0.123	0.242	0.322	0.454	0.568	0.342	0.909	1.179	1.088	1.469	1.640	1.257	0.942	0.897	0.917	0.849	0.809	0.883

TABLE VI. Comparison of MAE, RMSE, and R Values for SC-FITNET Imputation at Different Percentages of Missingness

Input	MAE (mm)						RMSE (mm)						R					
P	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG
P1	**0.072**	**0.133**	**0.159**	**0.238**	**0.299**	**0.180**	**0.812**	**1.074**	**0.918**	**1.262**	**1.409**	**1.095**	**0.953**	**0.917**	**0.940**	**0.885**	**0.851**	**0.909**
P2	0.081	0.153	0.189	0.277	0.347	0.209	0.873	1.187	1.034	1.362	1.529	1.197	0.946	0.896	0.921	0.864	0.826	0.890
P3	0.087	0.169	0.209	0.302	0.382	0.230	0.921	1.278	1.116	1.424	1.644	1.277	0.939	0.876	0.906	0.848	0.786	0.871
P4	0.093	0.176	0.219	0.318	0.401	0.241	0.956	1.302	1.128	1.463	1.686	1.307	0.935	0.873	0.906	0.840	0.774	0.866
P5	0.103	0.195	0.245	0.348	0.443	0.267	1.027	1.432	1.250	1.560	1.824	1.419	0.923	0.844	0.882	0.816	0.732	0.839
P6	0.100	0.192	0.240	0.342	0.432	0.261	1.002	1.397	1.220	1.537	1.771	1.385	0.928	0.851	0.887	0.822	0.747	0.847
P7	0.101	0.191	0.238	0.340	0.434	0.261	1.012	1.404	1.234	1.541	1.807	1.400	0.927	0.853	0.888	0.823	0.741	0.846
P8	0.155	0.301	0.402	0.559	0.707	0.425	1.175	1.656	1.541	1.871	2.174	1.683	0.915	0.834	0.871	0.811	0.729	0.832
P9	0.101	0.193	0.238	0.341	0.435	0.262	1.028	1.430	1.253	1.556	1.829	1.419	0.924	0.845	0.882	0.818	0.725	0.839
P10	0.085	0.176	0.207	0.304	0.396	0.234	0.927	1.371	1.182	1.598	1.878	1.391	0.938	0.860	0.897	0.807	0.712	0.843

TABLE VII. Comparison of MAE, RMSE, and R Values for LSTM Imputation at Different Percentages of Missingness

Input	MAE (mm)						RMSE (mm)						R					
P	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG
P1	0.081	0.162	0.191	0.255	0.338	0.205	1.015	1.585	1.401	1.538	1.907	1.489	0.928	0.812	0.857	0.825	0.704	0.825
P2	0.080	0.160	0.188	0.251	0.335	0.203	1.016	1.584	1.399	1.539	1.903	1.488	0.928	0.813	0.857	0.825	0.706	0.826
P3	0.082	0.163	0.194	0.258	0.345	0.209	1.017	1.584	1.401	1.536	1.909	1.489	0.927	0.813	0.857	0.825	0.704	0.825
P4	**0.080**	**0.159**	0.187	**0.248**	0.333	**0.201**	**1.010**	**1.581**	**1.398**	**1.532**	**1.900**	**1.484**	**0.928**	**0.813**	**0.857**	**0.827**	**0.708**	**0.827**
P5	0.083	0.165	0.195	0.261	0.346	0.210	1.018	1.587	1.401	1.544	1.909	1.492	0.927	0.812	0.857	0.823	0.704	0.825
P6	0.083	0.165	0.195	0.262	0.348	0.210	1.017	1.586	1.399	1.541	1.909	1.491	0.927	0.812	0.857	0.824	0.705	0.825
P7	0.084	0.169	0.203	0.270	0.359	0.217	1.014	1.582	1.399	1.538	1.907	1.488	0.928	0.813	0.857	0.825	0.705	0.826
P8	0.080	0.159	**0.185**	0.249	**0.332**	**0.201**	1.025	1.590	1.405	1.548	1.919	1.497	0.926	0.811	0.856	0.822	0.700	0.823
P9	0.082	0.164	0.195	0.261	0.347	0.210	1.018	1.587	1.401	1.547	1.913	1.493	0.927	0.812	0.857	0.823	0.703	0.824
P10	0.081	0.162	0.192	0.257	0.341	0.207	1.013	1.584	1.401	1.538	1.906	1.488	0.928	0.813	0.857	0.825	0.706	0.826

Note: The best results obtained are made bold.

A. Effect of Different Imputation Methods on Rainfall Data Series at Different Input P and Percentages of Missingness

Table IV, V, VI, and VII show the effects of different imputation methods on rainfall data series at different input p and missing rates. As seen in Table IV, the performances of FFNN increased as the input p decreased. Performance measures such as MAE, RMSE, and R show that FFNN achieved the best accuracy in total when input $p4$ was applied to the network. The average values of MAE, RMSE, and R measures for FFNN imputation were 0.299 mm, 1.251 mm, and 0.884 at $p4$, respectively.

From Table V, among the input p values, the first and second input ($p1$, $p2$) demonstrated good performances for predicting missing rainfall data. In particular, the SC-FFNN imputation for $p1$ showed excellent performance in estimating the various percentages of missingness in terms of MAE, RMSE, and R. The SC-FFNN imputation achieved an average accuracy of 90 %. The average MAE and RMSE measures of SC-FFNN were 0.183 mm and 1.142 mm at $p1$, respectively.

Meanwhile, the SC-FITNET imputation achieved optimal performance when the input $p1$ was used with an average accuracy of 90.9 %, as shown in Table VI. The average MAE and RMSE values are 0.180 mm 1.095 mm, respectively. On the other hand, performance measure such as MAE indicates the LSTM imputation achieved the lowest average error at the $p4$ and $p8$ as in Table VII. For the RMSE and R measures, the LSTM imputation obtained the best performances when input $p4$ was used, with an average value of 1.484 mm and 0.827, respectively. Overall, input $p1$ is the significant input for SC-FITNET and SC-FFNN imputation, while input $p4$ is the significant input for FFNN and LSTM imputation to achieve optimal imputation performances.

Furthermore, the study indicates the different missing rates would impact the accuracy of the missing data imputation. For example, when the missing rates increased from 10% to 50% at input $p1$, the MAE and RMSE measures increased from [0.072 mm, 0.831 mm] to [0.301 mm, 1.473 mm] respectively, but R decreased from 0.951 to 0.842 when using SC-FFNN imputation. Overall, the same input p that achieved the lowest mean absolute error (MAE) might also achieve the highest correlation coefficient (R). However, at 10% and 20% missingness, this study revealed that the same input p with the lowest value of MAE achieved the second-highest value of R instead of the highest value. This happens when the SC-FFNN imputation is able to measure the error between the predicted and the eventual outcomes accurately, but the R measures of correlation and dependence between the predicted and observed rainfall were statistically not the strongest.

A closer inspection revealed that the values of MAE for the four imputation methods linearly increased when the proportions of missing values increased. However, the values of RMSE linearly increased when the dataset had more than 30% missing values. This study supports the previous findings of Gill [51], Lee and Huber [52], Shang [53], Kim [54], and Ayilara [55] that the performance of imputation decreased when the proportion of missingness increased. According to Gill [51], the effect of missing data in information becomes very significant for hydrologic predictions as the percentage of missing data increases. Hence, this study concluded that more missing rainfall data in the dataset results in a poorer model performance, which is consistent with previous research [51]-[55].

B. Effect of Data Pre-processing Methods on Missing Data Prediction Performance

To investigate the effect of pre-processing data on the precision of missing rainfall imputation, a min-max normalization was used as a benchmark pre-processing data. The best performances obtained from the four models tabulated in Table IV, V, VI and VII were compared with the min-max normalization approach. For the min-max normalization approach, the raw meteorological and rainfall data were normalized as follows:

$$input\ p = \frac{X - min(X)}{max(X) - min(X)} \tag{15}$$

where $min(x)$ is the minimum value, $max(x)$ is the maximum value, and X is the data point.

Table VIII, IX and X show the effect of two different data pre-processing methods on the missing data estimation performance in terms of MAE, RMSE and R. For the min-max normalization approach, this study revealed that the LSTM imputation outperformed the other three models due to its capability to correlate the features in data. The performance measures such as MAE, RMSE and R show that the LSTM imputation achieved the lowest MAE and RMSE but highest R, as in Table VII, IX and X. The SC-FITNET was the second place with an average accuracy of 59%, followed by SC-FFNN, and FFNN imputation at an average accuracy of 55% and 49%, respectively. However, the performances of SC-FITNET, SC-FFNN and FFNN became unreliable as the percentage of missing data increases. The min-max normalization approach leads to inaccurate prediction due to the presence of zeros during the long dry periods. As a result, the three neural network models were not able to estimate the missing rainfall accurately. Hence, the min-max normalization approach is not suitable to be used for the long dry periods because it does not handle outliers very well.

TABLE VIII. Result on Imputation Process - Mean Absolute Error (MAE)

Missing rates	min-max				proposed work - PCA			
	FFNN	SC-FFNN	SC-FITNET	LSTM	FFNN	SC-FFNN	SC-FITNET	LSTM
10%	0.707	0.480	0.413	**0.107**	0.108	0.072	**0.072**	0.080
20%	1.418	0.968	0.833	**0.203**	0.212	0.134	**0.133**	0.159
30%	2.125	1.442	1.204	**0.248**	0.280	0.163	**0.159**	0.187
40%	2.846	1.938	1.617	**0.348**	0.397	0.244	**0.238**	0.248
50%	3.551	2.412	2.026	**0.450**	0.497	0.301	**0.299**	0.333
Avg	2.129	1.448	1.219	**0.271**	0.299	0.183	**0.180**	0.201

TABLE IX. Result on Imputation Process - Root Mean Square Error (RMSE)

Missing rates	min-max				proposed work - PCA			
	FFNN	SC-FFNN	SC-FITNET	LSTM	FFNN	SC-FFNN	SC-FITNET	LSTM
10%	3.101	2.198	1.838	**1.127**	0.871	0.831	**0.812**	1.010
20%	4.471	3.158	2.661	**1.659**	1.210	1.108	**1.074**	1.581
30%	5.301	3.724	2.956	**1.476**	1.103	0.952	**0.918**	1.398
40%	6.158	4.351	3.421	**1.716**	1.435	1.344	**1.262**	1.532
50%	6.894	4.866	3.850	**2.063**	1.638	1.473	**1.409**	1.900
Avg	5.185	3.659	2.945	**1.608**	1.251	1.142	**1.095**	1.484

TABLE X. Result on Imputation Process - Correlation Coefficient, R

Missing rates	min-max				proposed work - PCA			
	FFNN	SC-FFNN	SC-FITNET	LSTM	FFNN	SC-FFNN	SC-FITNET	LSTM
10%	0.693	0.754	0.794	**0.910**	0.947	0.951	**0.953**	0.928
20%	0.522	0.593	0.613	**0.792**	0.896	0.912	**0.917**	0.813
30%	0.495	0.558	0.612	**0.840**	0.914	0.936	**0.940**	0.857
40%	0.416	0.474	0.525	**0.776**	0.856	0.874	**0.885**	0.827
50%	0.343	0.409	0.409	**0.637**	0.808	0.842	**0.851**	0.708
Avg	0.494	0.558	0.591	**0.791**	0.884	0.903	**0.909**	0.827

Note: The best results obtained are made bold.

On the other hand, the four models achieved higher performances when the proposed PCA data pre-processing approach was used. It shows that the proposed significant input was able to help the four models to estimate the missing time series at higher accuracy compared to the min-max approach. Performance measures such as MAE and RMSE show that the SC-FITNET has the lowest error rates among the other three models, while the SC-FFNN imputation was in second place. Furthermore, the correlation and coefficient, R-value indicates the SC-FITNET scored the highest average accuracy of 90.9%, followed by SC-FFNN, FFNN and LSTM imputation. It shows that the adaptation of sine cosine algorithm into the existing neural network (SC-FITNET and SC-FFNN) was able to optimize the neural network and achieved higher accuracy but lower MAE and RMSE values compared to the FFNN imputation. The possible reason is that the position-updating function of SC-FITNET and SC-FFNN could positively optimize the entire search space for the best weights and biases of the neural networks and consequently increase the accuracy of the imputation.

Furthermore, the LSTM performed slightly better when the proposed PCA data pre-processing approach was used than the min-max approach. However, this study revealed that SC-FITNET and SC-FFNN slightly indicate a better prediction performance compared to the LSTM model. In addition to that, recent studies have shown that temporal convolutional networks (TCN) [56], and multilayer perceptron (MLP) [57] can outperform recurrent models such as LSTM. The LSTM model may require a large amount of data to perform better than the other methods. In terms of computational time, the LSTM model required more time to perform the missing data estimation process than the three models, FFNN, SC-FFNN and SC-FITNET (results not shown here). Hence, the FFNN, SC-FFNN and SC-FITNET models have the advantage of being computationally less costly compared to the LSTM model. In particular, there is a reduction of the average training time in the three models, approximately four times less than the LSTM model.

Overall, the SC-FITNET imputation has proven to be the top performer when the proposed PCA data pre-processing approach was used, while the LSTM imputation demonstrated the top performer for the min-max normalization approach.

VI. Conclusion

We investigated the potential of using meteorological and rainfall data from nearest neighbor gauging stations for infilling missing rainfall data. Before the imputation, this study introduced PCA to extract significant features from the meteorological data. The comparison of different combination input in imputation was presented and evaluated using four imputation methods, SC-FITNET, SC-FFNN, LSTM and FFNN. With medium size data of 11,680 real-life records, the four methods were trained and compared at five different percentages of missingness under MCAR conditions (10%, 20%, 30%, 40%, and 50%). The study concluded that the proposed SC-FITNET imputation has a higher capability in treating missing values for the PCA pre-processed dataset than than the LSTM, SC-FFNN and FFNN imputation in terms of MAE, RMSE, and R. By adopting the position-updating function, the proposed SC-FITNET imputation successfully achieved better accuracy in missing data estimation as compared to the other three approaches. Hence, the results of the proposed SC-FITNET imputation in this work support its use for infilling real-life missing rainfall data. In addition, the study revealed that the meteorological data (non-precipitation) and rainfall data (precipitation) from nearest neighbor stations are compatible and can be used as input for missing data imputation. The performances of the proposed PCA as data pre-processing have an obvious advantage over the benchmark.

For future work, considering a longer period of data, investigating other data pre-processing techniques and further testing the effectiveness of the proposed algorithm on different types of datasets are recommended. In addition, the imputed rainfall dataset could be used as an input in the hydrological data analysis. The imputed data could be employed to estimate the river flow and the occurrence of floods during the rainy season, to determine the severity and frequency of drought during the dry season, to design water supply, and other hydrological data analyses.

Acknowledgment

The authors would like to acknowledge the Malaysian Meteorological Department and Department of Irrigation and Drainage (DID), Sarawak, Malaysia, for providing the meteorological and rainfall data in this study. This work was supported/funded by the Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS/1/2018/ICT04/UTM/01/1). The authors sincerely thank Universiti Teknologi Malaysia (UTM) under Research University Grant Vot-20H04, Malaysia Research University Network (MRUN) Vot 4L876, and SLAI supported under Ministry of Higher Education Malaysia for the completion of the research. The work is partially supported by the SPEV project (ID: 2102-2021), Faculty of Informatics and Management, University of Hradec Kralove. We are also grateful for the support of Ph.D. students Michal Dobrovolny and Sebastien Mambou in consultations regarding application aspects from Hradec Kralove University, Czech Republic. The APC was funded by the SPEV project 2102/2021, Faculty of Informatics and Management, University of Hradec Kralove.

References

[1] P. Muñoz, J. Orellana-Alvear, P. Willems, and R. Célleri. "Flash-flood forecasting in an Andean mountain catchment—Development of a step-wise methodology based on the random forest algorithm," *Water*, vol. 10, no. 11, 2018, pp. 1519.

[2] S. Szewrański, J. Chruściński, J. Kazak, M. Świąder, K. Tokarczyk-Dorociak, and R. Żmuda, "Pluvial flood risk assessment tool (PFRA) for rainwater management and adaptation to climate change in newly urbanised areas," *Water*, vol. 10, no. 4, 2018, pp. 386.

[3] K.K. Kuok, S. Harun, S.M. Shamsuddin, and P.C. Chiu, "Evaluation of daily rainfall-runoff model using multilayer perceptron and particle swarm optimization feedforward neural networks," *Journal of Environmental Hydrology*, vol. 18, no. 10, 2010, pp. 1-16.

[4] N. Yang, B.H. Men, and C.K. Lin, "Impact analysis of climate change on water resources," *Procedia Engineering*, vol. 24, 2011, pp. 643-648.

[5] K.K. Kuok, S. Harun, and P.C. Chiu, "Hourly runoff forecast at different leadtime for a small watershed using artificial neural networks," *International Journal of Advances in Soft Computing and its Application*, vol. 3, 2011, pp. 68-86.

[6] R.A. Mcdonald, P.W. Thurston, and M.R. Nelson, "A Monte Carlo study of missing item methods," *Organizational Research Methods*, vol. 3, no. 1, 2000, pp. 71-92.

[7] P.E. McKnight, K.M. McKnight, S. Sidani, and A.J. Figueredo, "Missing data: A gentle introduction," Guilford Press. 2007.

[8] K.J. Lee and J.B. Carlin, "Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation," *American Journal of Epidemiology*, vol. 171, no. 5, 2010, pp. 624-632.

[9] Y. Gao, C. Merz, G. Lischeid, and M. Schneider, "A review on missing hydrological data processing," *Environmental earth sciences*, vol. 77, no. 2, 2018, pp. 47.

[10] S. Londhe, P. Dixit, S. Shah, and S. Narkhede, "Infilling of missing daily rainfall records using artificial neural network," *ISH Journal of Hydraulic Engineering*, vol. 21, no. 3, 2015, pp. 255-264.

[11] T. Canchala-Nastar, Y. Carvajal-Escobar, W. Alfonso-Morales, W.L. Cerón and E. Caicedo, "Estimation of missing data of monthly rainfall in southwestern Colombia using artificial neural networks, " *Data in brief*,

vol. 26, 2019, pp. 104517.
[12] P.C. Chiu, A. Selamat, O. Krejcar, and K.K. Kuok, "Missing rainfall data estimation using artificial neural network and nearest neighbor imputation," In Advancing Technology Industrialization Through Intelligent Software Methodologies, Tools and Techniques: Proceedings of the 18th International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT_19), IOS Press, vol. 318, 2019, pp. 132.
[13] M.R. Mispan, N.F.A. Rahman, M.F. Ali, K. Khalid, M.H.A. Bakar and S.H. Haron, "Missing river discharge data imputation approach using artificial neural network," Methodology, vol. 25, 2015, pp. 20.
[14] P.C. Chiu, A. Selamat and O. Krejcar, "Infilling missing rainfall and runoff data for Sarawak, Malaysia using gaussian mixture model based K-nearest neighbor Imputation," In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, Springer, 2019, pp. 27-38.
[15] W.Y. Lai, and K.K. Kuok, "A study on bayesian principal component analysis for addressing missing rainfall data," *Water Resources Management*, 2019, pp.1-14.
[16] S. Mirjalili, "SCA: A sine cosine algorithm for solving optimization problems," *Knowledge-Based Systems*, vol. 96, 2016, pp. 120-133.
[17] C. Qu, Z. Zeng, J. Dai, Z. Yi, and W. He, "A modified sine-cosine algorithm based on neighborhood search and greedy levy mutation," *Computational Intelligence and Neuroscience*, 2018.
[18] S. Das, A. Bhattacharya and A.K. Chakraborty, "Solution of short-term hydrothermal scheduling using sine cosine algorithm," *Soft Computing*, vol. 22, no. 19, 2018, pp. 6409-6427.
[19] S. Li, H. Fang, and X. Liu, "Parameter optimization of support vector regression based on sine cosine algorithm," *Expert Systems with Applications*, vol. 91, 2018, pp. 63-77.
[20] M.A. Tawhid, and P. Savsani, "Discrete sine-cosine algorithm (DSCA) with local search for solving traveling salesman problem," *Arabian Journal for Science and Engineering*, 2018, pp. 1-11.
[21] R.E. Chandler, V.S. Isham, N.A. Leith, P.J. Northrop, C.J. Onof, and H.S. Wheater, "Uncertainty in rainfall inputs," World Scientific/Imperial College Press, 2011.
[22] O. Stoner, and T. Economou, "An advanced hidden markov model for hourly rainfall time series," arXiv:1906.03846. 2019.
[23] T. Kashiwao, K. Nakayama, S. Ando, K. Ikeda, M. Lee and A. Bahadori, "A neural network-based local rainfall prediction system using meteorological data on the Internet: A case study using data from the Japan Meteorological Agency," *Applied Soft Computing*, vol. 56, 2017, pp. 317-330.
[24] M.H. Yen, D.W. Liu, Y.C. Hsin, C.E. Lin, and C.C. Chen, "Application of the deep learning for the prediction of rainfall in Southern Taiwan," *Scientific Reports*, vol. 9, no. 1, 2019, pp. 1-9.
[25] M. Chhetri, S. Kumar, P.P. Roy, and B.G. Kim, "Deep BLSTM-GRU model for monthly rainfall prediction: A case study of Simtokha, Bhutan," *Remote Sensing*, vol. 12, no. 19, 2020, pp.3174.
[26] S.K. Grange, and D.C. Carslaw, "Using meteorological normalisation to detect interventions in air quality time series," *Science of the Total Environment*, vol. 653, 2019, pp.578-588.
[27] L.I. Smith, "A tutorial on principal components analysis", 2002. Accessed: Jan. 3, 2020. [Online]. Available: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
[28] C. Skittides, and W.G. Früh, "Wind forecasting using principal component analysis," *Renewable Energy*, vol. 69, 2014, pp. 365-374.
[29] M. Hubert, P.J. Rousseeuw, and W. Van den Bossche, "MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers," *Technometrics*, vol. 61, no.4, 2019, pp. 459-473.
[30] Z. Zuśka, J. Kopcińska, E. Dacewicz, B. Skowera, J. Wojkowski, and A. Ziernicka–Wojtaszek, "Application of the principal component analysis (PCA) method to assess the impact of meteorological elements on concentrations of particulate matter (PM10): A case study of the mountain valley (the Sącz Basin, Poland)," *Sustainability*, vol. 11, no. 23, 2019, pp. 6740.
[31] Y.Y Choi, H. Shon, Y.J. Byon, D.K. Kim, S. Kang, "Enhanced application of principal component analysis in machine learning for imputation of missing traffic data," *Applied Science*, vol. 9, no. 10, 2019, pp. 2149.
[32] B.S. Harish, and S.V.A Kumar. "Anomaly based intrusion detection using modified fuzzy clustering," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol 4, no. 6, 2017, pp. 54-59, doi: 10.9781/ijimai.2017.05.002.
[33] T. Kurita, "Principal component analysis (PCA)," In: Ikeuchi K. (eds) *Computer Vision: A Reference Guide*, Springer, 2014.
[34] K. Pearson, "Principal components analysis," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 6, no.2, 1901, pp. 559.
[35] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, 1933, pp. 417-441.
[36] R. Khattree, and D.N. Naik "Multivariate data reduction and discrimination with SAS software," SAS Institute, 2000.
[37] G. Bebis, and M. Georgiopoulos, "Feedforward neural networks," *IEEE Potentials*, vol. 13, no. 4, 1994, pp. 27-31.
[38] S. Mirjalili, "How effective is the Grey Wolf optimizer in training multi-layer perceptrons," *Applied Intelligence*, vol. 43, no.1, 2015, pp.150-161.
[39] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," *Neural Computing*, vol. 9, no. 8, 1997, pp. 1735-1780.
[40] E. Mussumeci, and F.C. Coelho, "Large-scale multivariate forecasting models for Dengue-LSTM versus random forest regression," *Spatial and Spatio-temporal Epidemiology*, vol. 35, 2020, pp. 100372.
[41] S. Maya, and U. Ken, "DADIL: Data augmentation for domain-invariant learning," *Data Science and Pattern Recognition*, vol. 4, no. 2, 2020, pp. 33–49.
[42] J.C.W. Lin, Y. Shao, Y. Djenouri and U. Yun, "ASRNN: A recurrent neural network with an attention model for sequence labeling," *Knowledge-Based Systems*, vol. 212, 2020, pp. 106548.
[43] Hourly meteorological dataset for Kuching station: 2002 to 2003, Malaysian Meteorological Department, Selangor, Malaysia, October 2019.
[44] Hourly rainfall datasets for Sungai Merang station and nearest neighbor stations: 2002 to 2003, Department of Irrigation and Drainage (DID), Sarawak, Malaysia, October 2019.
[45] A.J. Henry, N.D. Hevelone, S. Lipsitz, and L.L. Nguyen, "Comparative methods for handling missing data in large databases," *Journal of Vascular Surgery*, vol. 58, no. 5, 2013, pp. 1353-1359.
[46] J.R. Cheema, "Some general guidelines for choosing missing data handling methods in educational research," *Journal of Modern Applied Statistical Methods*, vol. 13, no. 2, 2014, pp. 3.
[47] P. Zhu, Q. Xu, Q. Hu, C. Zhang, and H. Zhao, "Multi-label feature selection with missing labels," *Pattern Recognition*, vol. 74, 2018, pp. 488-502.
[48] H. Hassani, M. Kalantari, and Z. Ghodsi, "Evaluating the performance of multiple imputation methods for handling missing values in time series data: A study focused on East Africa, soil-carbonate-stable isotope data," *Stats*, vol. 2, no. 4, 2019, pp. 457-467.
[49] S. Oba, M.A. Sato, I. Takemasa, M. Monden, K.I. Matsubara, and S. Ishii, "A Bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, no.16, 2003, pp. 2088-2096.
[50] R.J. Little, and D.B. Rubin, "Statistical analysis with missing data," John Wiley & Sons, 2014.
[51] M.K. Gill, T. Asefa, Y. Kaheil, and M. McKee, "Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique," *Water Resources Research*, vol. 43, no.7, 2007.
[52] J. H. Lee, and Jr.J. Huber, "Multiple imputation with large proportions of missing data: How much is too much?," In United Kingdom Stata Users' Group Meetings 2011 (No. 23). Stata Users Group, 2011.
[53] Q. Shang, Z. Yang, S. Gao, and D. Tan, "An imputation method for missing traffic data based on FCM optimized by PSO-SVR," *Journal of Advanced Transportation*, 2018.
[54] T. Kim, W. Ko, and J. Kim, "Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting," *Applied Sciences*, vol. 9, no. 1, 2019, pp. 204.
[55] O.F. Ayilara, L. Zhang, T.T. Sajobi, R. Sawatzky, E. Bohm, and L.M. Lix, "Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry," *Health and Quality of Life Outcomes*, vol. 17, no. 1, 2019, pp. 106.
[56] S. Bai, J.Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint arXiv:1803.01271, 2018.

[57] A. Cecaj, M. Lippi, M. Mamei, and F. Zambonelli, "Comparing deep learning and statistical methods in forecasting crowd distribution from Aggregated Mobile Phone Data," *Applied Sciences*, vol. 10, no. 18, 2020, pp. 6580.

Po Chan Chiu

Po Chan Chiu is currently pursuing Ph.D degree in Computer Science from Universiti Teknologi Malaysia (UTM). She received the M.Sc. in information technology from the Universiti Malaysia Sarawak (UNIMAS), in 2010. She was the Software Engineer at private companies for 3 years. Her research interests include artificial intelligence, optimization, data analytics and neural networks.

Ali Selamat

Ali Selamat is currently a Full Professor with Universiti Teknologi Malaysia (UTM), Malaysia. He has also been the Dean of the Malaysia Japan International Institute of Technology (MJIIT), UTM, since 2018. An academic institution established under the cooperation of the Japanese International Cooperation Agency (JICA) and the Ministry of Education Malaysia (MOE) to provide the Japanese style of education in Malaysia. He is also a Professor with the Software Engineering Department, School of Computing, UTM and the Chair of the IEEE Computer Society Malaysia Section. He has published more than 120 research articles with IF JCR, with more than 2400 citations received in the Web of Science and h-index 26. His research interests include software engineering, software process improvement, software agents, web engineering, information retrievals, pattern recognition, genetic algorithms, neural networks, soft computing, collective computational intelligence, strategic management, key performance indicator, and knowledge management. He is on the Editorial Board of the journal Knowledge-Based Systems (Elsevier).

Ondrej Krejcar

Ondrej Krejcar is currently a Full Professor of systems engineering and informatics with the University of Hradec Kralove, Czech Republic. He is also the Vice-Dean for science and research at the Faculty of Informatics and Management, UHK. He is also the Director of the Center for Basic and Applied Research, University of Hradec Kralove. At the University of Hradec Kralove, he is a Guarantee of the Doctoral Study Programme in Applied Informatics, where he is focusing on lecturing on smart approaches to the development of information systems and applications in ubiquitous computing environments. His h-index is 20 (according Web of Science), with more than 1500 citations received in the Web of Science. He has published more than 110 research articls with IF JCR. He has a number of collaborations throughout the world (e.g., Malaysia, Spain, U.K., Ireland, Ethiopia, Latvia, and Brazil). His research interests include control systems, smart sensors, ubiquitous computing, manufacturing, wireless technology, portable devices, biomedicine, image segmentation and recognition, biometrics, technical cybernetics, and ubiquitous computing. His second area of interest is in biomedicine (image analysis), as well as biotelemetric system architecture (portable device architecture and wireless biosensors), and the development of applications for mobile devices with use of remote or embedded biomedical sensors. Dr. Krejcar has also been a Management Committee Member substitute of the project COST CA16226, since 2017. In 2018, he was the 14th Top-Peer Reviewer in Multidisciplinary in the World according to Publons. He is on the Editorial Board of Sensors (MDPI) with JCR Index and several other ESCI indexed journals. He has been the Vice-Leader and a Management Committee Member at WG4 of the project COST CA17136, since 2018. Since 2019, he has been the Chairman of the Program Committee of the KAPPA Program, Technological Agency of the Czech Republic, as a Regulator of the EEA/Norwegian Financial Mechanism in the Czech Republic (2019–2024). Since 2014, he has been the Deputy Chairman of the Panel 7 (Processing Industry, Robotics and Electrical Engineering) of the Epsilon Program, Technological Agency of the Czech Republic.

King Kuok Kuok

King Kuok Kuok is a senior lecturer at Swinburne University of Technology Sarawak Campus. He received his MEng from the UNIMAS in 2004 and Ph.D. from the UTM in 2010. He was the Field Engineer for Hydrological and Water Resources Branch, Department of Irrigation and Drainage, State of Sarawak, Malaysia from 2002 to 2009 and the Road, Civil and Structural Design Engineer at private companies for more than 10 years. His research interests include water resources, water supply, hydrology, artificial intelligence and building information modeling.

Enrique Herrera-Viedma

Enrique Herrera-Viedma is a Professor of Computer Science and the Vice-President of Research and Knowledge Transfer with the University of Granada. His H-index is 85 with more than 25000 citations received in Web of Science and 97 in Google Scholar with more than 38500 citations received. His current research interests include group decision-making, consensus models, linguistic modeling, aggregation of information, information retrieval, bibliometric, digital libraries, Web quality evaluation, recommender systems, and social media. He has been identified as one of the World's Most Influential Researchers by Shanghai Center and Thomson Reuters/Clarivate Analytics in both the computer science and scientific engineering categories in 2014–2020. Prof. Herrera-Viedma was the 2019–2020 Vice-President of Publications with the IEEE SMC Society and an Associate Editor of several journals, such as the IEEE Transactions on Fuzzy Systems, the IEEE Transactions on Systems, Man, and Cybernetics: Systems, Information Sciences, Applied Soft Computing, Soft Computing, Fuzzy Optimization and Decision Making, and Knowledge-Based Systems.

Giuseppe Fenza

Giuseppe Fenza received the Ph.D. degree in Computer Sciences at the University of Salerno, Italy, in 2009. From 2009 until now, he collaborates to several research initiatives mainly focused on Knowledge Extraction from unstructured resources defining intelligent systems based on the combination of techniques from Soft Computing, Semantic Web, areas in which he has many publications. He has been deeply involved in several EU and Italian Research and Development projects on ICT and, in particular, on Situation Awareness, Service Discovery, Enterprise Information Management and e-Commerce. He serves as Associate Editor in international journals, such as: Neurocomputing, International Journal of Grid and Utility Computing, International Journal of Engineering Business Management. He has published extensively about: Fuzzy Decision Making, Ontology Elicitation, Situation and Context-Awareness, Semantic Information Retrieval. Recently, he is working in the field of Big Data, Social Media Analytics, and Web Intelligence by proposing novel methods for instance, to support microblog summarization, time-aware information retrieval and recommendation extraction. He is currently an Assistant Professor in Computer Science at the Department of Management and Innovation Systems, University of Salerno, Italy.

View publication stats