VOL. 10, NO 22, DECEMBER, 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

MISSING RIVER DISCHARGE DATA IMPUTATION APPROACH USING ARTIFICIAL NEURAL NETWORK

M. R.Mispan[1], N. F. A. Rahman[2], M. F. Ali[2], K. Khalid[3], M. H. A. Bakar[4] and S. H. Haron[5]
[1]Soil and Water Management MARDI, Serdang, Selangor, Malaysia
[2]Faculty of Civil Engineering, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
[3]Faculty of Civil Engineering, Universiti Teknologi MARA Pahang, Jengka, Pahang, Malaysia
[4]Center for Research and Innovation, UniKL Malaysian Spanish Institute, Kulim, Kedah, Malaysia
[5]Faculty of Science and Technology, Universiti Kebangsaan Malaysia UKM, Bangi, Selangor, Malaysia
E-Mail: radzali@mardi.gov.my

ABSTRACT

The issue with missing data in hydrological models are very common and it occurs when no data value was stored during observation. In modelling, the missing data can affect the conclusion that can be drawn from the dataset. This paper presents the study on Levenberg-Marquadt back propagation algorithm in predicting missing stream flow data in Langat River Basin. Data series from the upper part of Langat River Basin were applied to build the Artificial Neural Network model. The result indicated good performance of the model with Pearson Correlation, r = 0.97261 for training data and overall data shows r = 0.91925. The study reveals that Levenberg-Marquadt back propagation algorithm for ANN can simulate well in the daily missing stream flow prediction if the model customizes with good configuration.

Keywords: artificial neural networks, hydrological modelling, missing data.

INTRODUCTION

Over the years, although there are advances in missing data imputation methods, the issues revolving missing data remain largely unsolved. Numerous techniques have been developed for data imputation and it can be classified into two classes: statistical methods and computer intelligence methods. Due to the common practise use of statistical methods in handling missing data problems, the shift for computational intelligence methods to gain significant consideration despite the higher accuracy may takes some time. The merits of both these classes have been discussed at length in the literature and the pro and cons for each have been identified [1, 2].

There are several reasons why data may be missing. They may be missing because equipment malfunctioned, the weather was terrible, people got sick, or the data were not entered correctly. Why the data is missing is not as important compared to the value of its observation.

In modelling climate changes study for upper part of Langat River Basin using Soil and Water Assessment Tool (SWAT), a semi distributed hydrological model, the calibration and validation of SWAT model becomes difficult die to the absence of missing data of river discharge at Sg Langat. The numbers of missing data are varies from day to month to years. Thus it will raise the bias factors of the hydrological model itself. In managing the missing data, artificial neural networks (ANNs) has been adopted as an approach to solve the problem before the SWAT model evaluation been carried out.

Artificial neural networks (ANNs) have become one of the most promising tools for missing data imputation. In most studies, ANN has been demonstrated to show superior result compared to the traditional statistical approaches. The model are relationships between input and output data without detailed knowledge of the processes under investigation, by finding an optimum set of network parameters through the training process [3].

This paper addresses the issue of understanding of a trained ANN model by using Levenberg-Marquardt back propagation. Then the applicability of the algorithm within the watershed is been carried out. The objective of this research is to analyse the potential of Levenberg-Marquardt back propagation of Artificial Neural Network in predicting missing stream flow data. This research develop ANN model to predict infilling missing streamflow data for Langat River Basin for calibration process using a hydrologic model, Soil and Water Assessment Tool.

Study area and data availability

Langat River basin is chosen as a study area due to its water quality declination and also the availability of data. Role of Langat River basin are providing raw water supply to approximately 1.3 million people in the area[4, 5] To predict missing river discharge value, rainfall data needed as the input to the nonlinear hydrological modelling using artificial neural networks approach [6, 7]. The list of the stations available in upper part of Langat River Basin is as shown in Table-1.

Table-1. Location of rainfall station.

Station Id	Station name	Station type
2917001	RTM Kajang	Daily
3018107	Ldg. Dominion	Daily
3118069	Pemasokan Ampang	Daily
3118102	Sek.Keb.Kg.Sg. Lui	Daily
3218101	TNB Pansoon	Daily
3119002	Ladang Sg Lui	Daily
3119104	Bt.30, Jalan Genting Peres	Daily

The rainfall stations are distributed on the upper part of Langat catchment as shown in Figure-1.

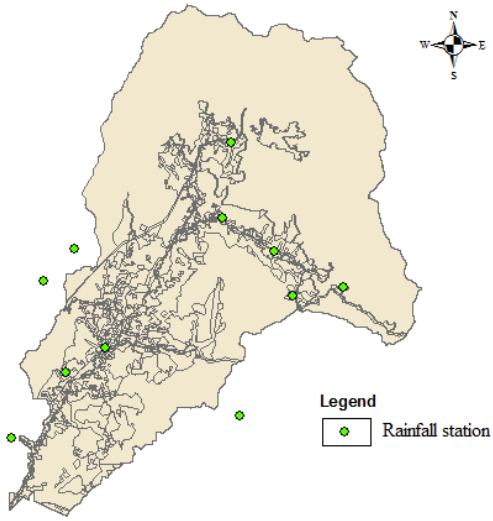

Figure-1. Rainfall stations location of Langat River Basin.

Data sets of 2 years were used for this study. For model set up, 70 percent of data from 1980 were used. For test and validation, 15% for test and another 15% data of 1980 were used for validation. Then the net was tested on year 1981 to see values of missing data. 1980 was chosen as the baseline as that year has complete river discharge data. It means zero missing data.

The DID (Department of Irrigation) provided river discharge at stream gauging stations as shown in Table-2.

Table-2. Location of river discharge.

ID	Station information			
	Name	**Type**	**Latitude** (N)	**Longitude** (E)
3118445	Kajang	Daily	030 10’ 25”	101052’ 20”

METHODOLOGY

Artificial Neural Network is a simple predictive algorithm tries to mimic the relationship between the input and the output variables. The function derived in such routines is a direct linear or non-linear function between input and output variables. It is truly said that the working of ANN takes its roots from the neural network residing in human brain[8]. ANN operates on something referred to as Hidden State. These hidden states are similar to neurons. Each of this hidden state is a transient form which has a probabilistic behaviour.

Neural networks, with their outstanding ability to derive meaning from problematical or imprecise data, can be used to extract patterns and detect tendencies that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to scrutinise. This expert can then be used to provide projections given new situations of interest.

Different points of interest include

Adaptive learning: An ability to learn how to do tasks based on the data given for exercise or initial practice.

Self-Organisation: An ANN can create its own organisation or depiction of the information it obtains during exercise time.

Real time operation: ANN computations may be carried out in parallel, and special hardware devices are

being designed and manufactured which take advantage of this capability.

Fault tolerance via redundant information coding: Partial destruction of a network leads to the corresponding degradation of performance. However, some network competences may be retained even with major network impairment.

Neural networks take a dissimilar approach to problem solving than that of conventional computers. Conventional computers use an algorithmic method i.e. the computer follows a set of commands in order to solve a problem. If the specific steps that the computer needs to follow are known the computer cannot solve the problem. That restricts the problem solving ability of conventional computers to problems that we already comprehend and know how to solve. But computers would be so much more valuable if they could do things that we don't exactly know how to do [8].

Neural systems process data in a comparative manner the human mind does. The system is balanced of a substantial number of exceptionally interconnected preparing essentials (neuron) working in parallel to take care of a particular issue. Neural systems learn by sample [9]. They can't be customized to perform a definite assignment. The samples must be chosen precisely generally valuable time is squandered or surprisingly more dreadful the system may be working incorrectly.

Then again, ordinary PCs utilize an intellectual way to deal with critical thinking; the way the issue is to comprehended must be known and expressed in little unmistakable directions. These guidelines are then changed over to a high-level language program and afterward into machine code that the computer can fathom [10, 11]. These machines are absolutely predictable; if anything turns out badly is because of a product or equipment slip.

Neural networks and conventional algorithmic computers are not in rivalry but accompaniment for each other. There are tasks are more suited to an algorithmic approach like arithmetic operations and tasks that are more suited to neural networks. Even more, a large number of tasks, require systems that use a mixture of the two approaches (normally a conventional computer is used to supervise the neural network) in order to execute at full efficiency.

A proper selection of input variables is crucial in building ANN models. The parameters were: 1) input data, 2) algorithm, 3) number of hidden neurons in hidden layer, and 4) learning [12].

The NARX model is based on the linear ARX model, which is commonly used in time-series modelling [11]. The defining equation for the NARX model is shown in equation 1[13].

$$y(t)=f(y(t-1),y(t-2),\ldots,y(t-ny),u(t-1),u(t-2),\ldots,u(t-nu)) \quad (1)$$

where the next value of the stream flow y(t) is regressed on previous values of the output streamflow and previous values of an independent (exogenous) rainfall. Figure-2 shows the neural network NARX framework for infilling stream flows missing data ANN model. To model infilling missing data ANN for stream flow, the rainfall stations that scattered around the upper part of Langat River Basin were evaluated to see the representative of each rainfall for the watershed [5, 6]. Only four stations from 9 rainfall station scattered in the watershed were chosen as input to the ANN model to reduce model complexity.

The four stations were chosen based on their location and data availability and reliability for the watershed. Input layer consist of four rainfall station scattered at the upper part of Langat river basin, the hidden layer chosen is 10 and the output layer is the stream flow located at Kajang.

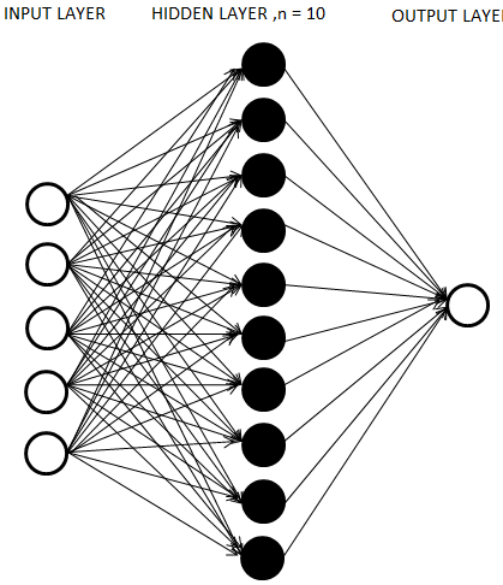

Figure-2. NARX framework.

RESULTS AND INFILLING MISSING DATA ANALYSIS

Model training is the initial testing of the model train the dataset to get input output relationship. The Pearson Correlation (r) was used to evaluate the model performance. R value ranges from 0-1, shows the correlation between the observed versus the simulated values. If the R value is less than or very close to zero, the model performance is considered unacceptable or poor. In contrast, if the values are equal to one, then the model prediction is considered perfect. For this project, training results were considered very good with 70 % of data trained resulted in Pearson Correlation, r = 0.97261. 255 days in 1980 were involved in the training stage. After the

training stage, the dataset was tested for validation and test stage and both stages show Pearson Correlation above 0.9 values. The algorithm was tested on the year 1980 and the result shows a very good r value with 0.91925 for the whole dataset. Figure-2 shows the stream flow dataset distribution for training stage and overall analysis on the data set. The simulated stream flow was significantly fit with the observed flow data pattern as shown in Figure-3.

With r values ranging above 0.9 for all stages, the performance of Levenberg Marquadt algorithm to predict infilling daily missing stream flow data in upper part of Langat River basin was satisfied.

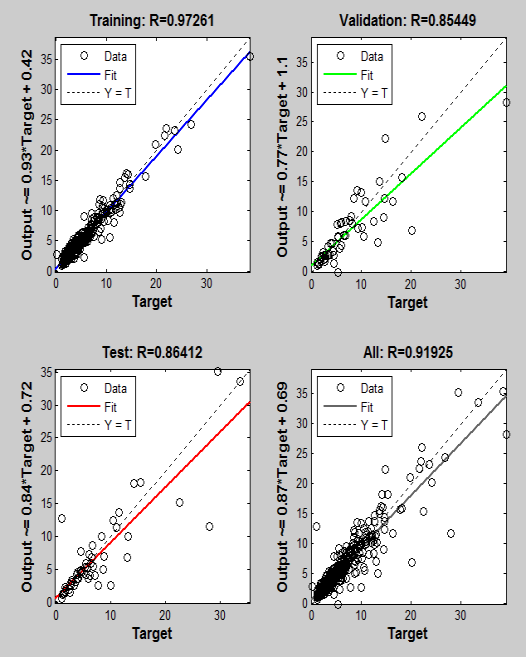

Figure-3. R value for all stages of simulation.

Figure-4 shows the observed river discharge at Kajang discharge satation versus river discharge simulated using the Lavenberg Marquadt algorithm.

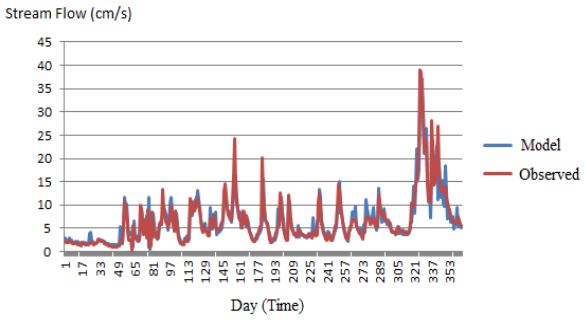

Figure-4. Hydrograph.

Based on the hydrograpgh, we can observed that the pattern of daily discharge if almost fitted with each other. The coefficient of determination (R^2) was used to evaluate the model performance. R^2 value ranges from 0-1, shows the correlation between the observed versus the simulated values. If the R^2 is less than or very close to zero, the model performance is considered unacceptable or poor. In contrast, if the values are equal to one, then the model prediction is considered perfect [6, 7]. The calculation of R^2 is using Equation (2). The result of R^2 is shown in Figure-5.#

#

$$r^2 = \left(\frac{\sum_{i=1}^{n} (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^{n} (O_i - \bar{O})^2} \sqrt{\sum_{i=1}^{n} (P_i - \bar{P})^2}} \right)^2 \quad \text{##################(2)#}$$

#

#

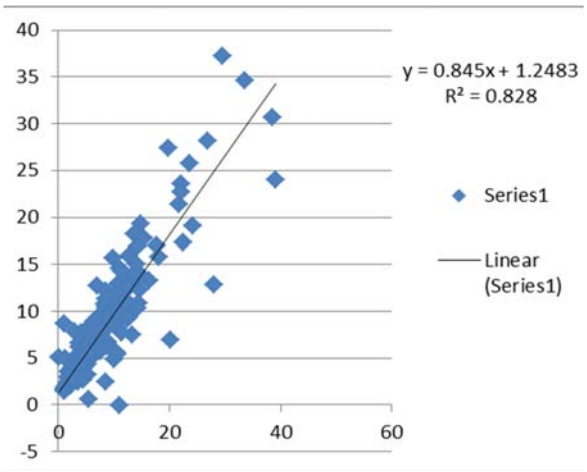

Figure-5. Coefficient of determination of simulated vs observed river discharge.

The net produce a very good result with R^2 = 0.828 as shown in Figure-5. After the net has been tested and the condition of the net is satisfied, the model was then run on 365 days of 1981 to get the daily missing data for year 1981.

Figure-6 and Figure-7 show the graph on the constructed daily missing data for the year 1981. In the year 1981, only the months of March and July had daily missing data problem. By using the Lavenberg Marquadt ANN model, the value for daily missing data had successfully been modelled. Thus, data gaps for continues hydrological modelling analysis can be reduced.

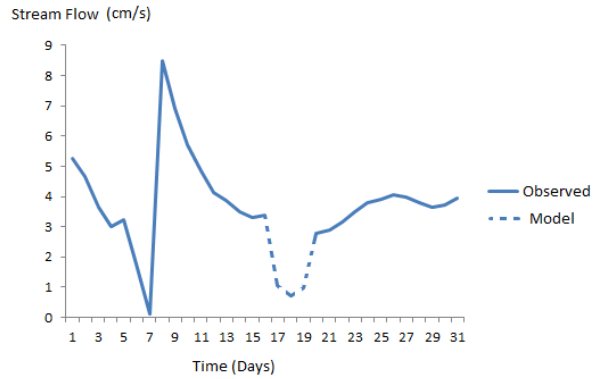

Figure-6. Reconstructed daily missing value for March 1981.

Based on Figure-6, there are four (4) days of missing data on the month of March 1981. The dates of missing data are starting from 16 to 19 March 1981.

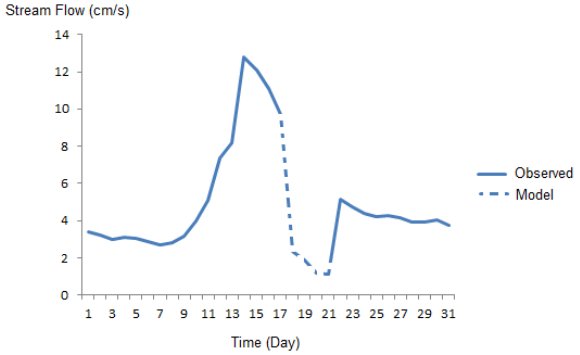

Figure-7. Reconstructed daily missing value for July 1981.

With reference to Figure-7, there are five (5) days of missing data on the month of July 1981. The dates of missing data are starting from 17 to 21 July 1981. From this research, we are able to retrieve the missing value using the ANN approach. Table-3 shows all the missing values for the year 1981 a represented as the dotted line in both Figure 6 and 7.

Table-3. Missing data imputation value.

Date of missing data	Imputation values by ANN
16 March 1981	3.394236702
17 March 1981	1.054353554
18 March 1981	0.732512518
19 March 1981	0.982702698
17 July 1981	9.694178
18 July 1981	2.321946
19 July 1981	1.891125
20 July 1981	1.169674
21 July 1981	1.139812

Based on the result, the study proved that Levenberg-Marquadt back propagation algorithm for ANN can simulate well in predicting the missing stream flow daily data if the model customize with good configuration. The finding in this research will help set up a good platform for running the simulation of SWAT model in upperpart of Langat River Basin for climate change impact studies. The net will be basis on infilling missing river discharge data on Sungai Langat at Kajang station for 30 years that will be discussed on another paper.

CONCLUSIONS

In this research, stream flow discharge of the upper part of Langat Catchment was successfully modeled in MATLAB environment. Based on this research, the daily precipitation missing data in the year 1981 has successfully been modelled from the baseline period 1980. From this study, the results obtained in training and validation are con sidered good which r values range from 0.91 to 0.97 for flow parameters.

Despite the disadvantages of ANN known as black box model, ANN can simulate the pattern of hydrological changes and imitate the actual river flow in Langat River Basin in the year 1980. Based on the pattern recognition of input and output, ANN can model the nonlinear hydrological processes with no spatial information and produce great results.

This ANN model-infilling techniques that can be rapidly deployed across large numbers of records and be highly beneficial in improving the data consistency. A further direction of this study is to simulate the stream flow for in filling missing data until the year 2010 for enhancement on hydrological model calibration.

ACKNOWLEDGEMENT

The project was supported by the Fundamental Research Grant Scheme (FRGS), Ministry of Education, Universiti Teknologi MARA (UiTM), Malaysia and SWAT Networks of Malaysia.

REFERENCES

[1] A.J. M. Jamil. 2012. Partial Least Squares Structural Equation Modelling with incomplete data; An investigation of the impact of imputation methods.

[2] A. Mohammad Kalteh. 2007. Rainfall-Runoff Modelling Using Artificial Neural Networks (ANNs).

[3] M. Kumar, N. S. Raghuwanshi, R. Singh, W. W. Wallender and W. O. Pruitt. 2002. Estimating Evapotranspiration using Artificial Neural Network. no. August, pp. 224-233.

[4] K. Khalid, M.F. Ali and N.F. Abd Rahman. 2014. The development and application of Malaysian Soil Taxonomy in SWAT Watershed Model. ISFRAM 2014. Proceedings of International Symposium on Flood Research and Management. pp. 79-88.

[5] M.F. Ali, N.F. Abd Rahman, K. Khalid. 2014. Discharge assessment by using integrated hydrologic model for environmental technology development. Journal of Advanced Materials Research. 911: 378-382.

[6] M. F. Ali, N. F. A. Rahman, K. Khalid and N. D. Liem. 2014. Langat River Basin Hydrologic Model Using Integrated GIS and ArcSWAT Interface. Appl. Mech. Mater. 567: 86-91.

[7] A.Y. Shamseldin. 1997. Application of neural network technique to rainfall-runoff modelling, J. Hydrol. 199(3-4): 272-294.

[8] J. Olsson. 2004. Neural networks for rainfall forecasting by atmospheric downscaling, J. Hydrol. Eng. 9(1): 1-12.

[9] 2000a. ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, Artificial neural networks in hydrology. I: Preliminary concepts, J. Hydrol. Eng. 5(2): 115-123.

[10] 2000b. ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, Artificial neural networks in hydrology. II: Hydrologic applications, J. Hydrol. Eng. 5(2): 124-137.

[11] Information on http://www.mathworks.com/help/nnet/ug/design-time-series-narx-feedback-neural-networks.html.

[12] C.M. Zealand, D.H. Burn and S.P. Simonovic. 1999. Short term streamflow forecasting using artificial neural networks. J. Hydrol. 214(1-4): 32-48.

[13] P. Coulibaly, F. Anctil and B. Bobe'e. 2000. Daily reservoir inflow forecasting using artificial neural networks with stopped training approach, J. Hydrol. 230(3-4): 244-257.