CORE

Metadata, citation and similar papers at core.ac.uk

Provided by Research Repository

Murdoch
UNIVERSITY

MURDOCH RESEARCH REPOSITORY

This is the author's final version of the work, as accepted for publication following peer review but without the publisher's layout or pagination. The definitive version is available at www.springerlink.com

Kajornrit, J., Wong, K.W. and Fung, C.C. (2012) *Estimation of missing precipitation records using modular artificial neural networks.* In: Huang, T.; Zeng, Z.; Li, C.; Leung, C.S. (Eds.) Neural Information Processing: Proceedings of the 19th International Conference on Neural Information Processing (ICONIP2012), Part 5, Lecture Notes in Computer Science, Vol 7667

http://researchrepository.murdoch.edu.au/11392/

Copyright © Springer-Verlag GmbH Berlin Heidelberg

It is posted here for your personal use. No further distribution is permitted.

Estimation of Missing Precipitation Records using Modular Artificial Neural Networks

Jesada Kajornrit, Kok Wai Wong, Chun Che Fung

School of Information Technology, Murdoch University
South Street, Murdoch, Western Australia, 6150
j_kajornrit@hotmail.com, {k.wong, l.fung}@murdoch.edu.au

Abstract. Estimation of missing precipitation records is one of the important tasks in hydrological study. The completeness of precipitation data leads to more accurate results from the hydrological models. This study proposes the use of modular artificial neural networks to estimate missing monthly rainfall data in the northeast region of Thailand. The simultaneous rainfall data from neighboring control stations are used to estimate missing rainfall data at the target station. The proposed method uses two artificial neural networks to learn the generalized relationship of rainfall recorded in dry and wet periods. Inverse distance weighting method and optimized weight of subspace reconstruction method are used to aggregate the final estimation value from both networks. The experimental results showed that modular artificial neural networks provided a higher accuracy than single artificial neural network and other conventional methods in terms of mean absolute error.

Keywords: Missing precipitation records, Modular artificial neural networks, Northeast region of Thailand, Inverse distance weighting method, Optimized weight of subspace reconstruction method.

1 Introduction

Precipitation data are one of the most important variables used in hydrological modeling in the assessment of streamflow and rainfall-runoff. These models fundamentally require the complete and reliable rainfall data records [1]. Normally, ground-based observations are the primary sources of rainfall data. A large number of rain gauge stations are installed throughout the study area to record the rainfall. However, in practice, rainfall records often contain missing data values due to malfunctioning of the equipment and/or other conditions. Such imperfect rainfall record could affect the performance of the hydrological models. Therefore, estimating missing rainfall data is an important task in hydrological modeling [2]. This study proposes the use of Modular Artificial Neural Networks (MANN) to estimate missing monthly rainfall data. This paper is organized as follows: Section 2 describes some of the related works. Section 3 illustrates four case studies and the dataset being used. Section 4 describes the details of MANN used in this study. Section 5 shows the experimental results and an analysis of the outcomes. Finally, a conclusion is presented in Section 6.

2 Related Works

In the last decade, many studies have been dedicated to address the missing rainfall data problem. Teegavarapu et al. [2] examined Inverse Distance Weighting Method (IDWM) and its variants to estimate the missing precipitation data. They suggested several ways to improve IDWM by defining some parameters and surrogate measures for distance used in IDWM. They concluded that using correlation coefficient as weight for revised IDWM and Artificial Neural Network (ANN) yielded better accuracy. Later, Teegavarapu et al. [3] improved Ordinary Kriging (OK) by using ANN to create semivariogram instead of using a prior definition of a mathematical function. This revised technique was used to estimate the missing precipitation data. The results showed that the use of ANN with OK had more advantages than the original OK. Nevertheless, Teegavarapu et al. [4] purposed a fixed functional set genetic algorithm method to derive the optimal functional forms for estimating the missing precipitation data. The method used genetic algorithm and non-linear optimization formulation to obtain functional form and its coefficients. The proposed method was compared with IDWM and Correlation Coefficient Weighting Method (CCWM). Their method showed improvement to IDWM and CCWM in term of root mean square error. Kim et al. [1] applied Regression Tree (RT) and ANN to construct missing precipitation data. Regression tree was used to create the list of influenced stations. These stations were then used to estimate the missing precipitation data by ANN models. The result showed that the use of RT + ANN provided better estimation than the use of RT or ANN alone. Piazza et al. [5] compared various spatial interpolation methods to create a serially complete monthly precipitation time series. Their study suggested that the best estimation result could be derived from the use of a combination method called residual kriging in which the residual from linear regression are interpolated by ordinary kriging method. Another comparison work is Kajornrit et al. [6]. They compared several spatial interpolation methods to estimate missing rainfall data in the northeast region of Thailand. They suggested the use of statistics of dataset as a guideline to select the appropriate estimation techniques. All works mentioned above used the single model to estimate missing rainfall data. Since the nature of rainfall data could be grouped into dry and wet period, the use of modular models may improve the estimation accuracy. Therefore, this study proposes the use of modular artificial neural networks to perform this task.

3 Four Case Studies and Dataset

The case study area selected sites in the northeast region of Thailand as illustrated in Figure 1. In this study, four rainfall stations are assumed to have missing rainfall data records (*target station*). The simultaneous rainfall data from neighboring stations (*control station*) are used to estimate the missing data at target station. Many researchers have recommended the use of three or four closest stations for application of IDWM [2]. This suggestion related to the work of Eischeid [7], which showed that inclusion of more than four stations does not significantly improve the interpolation and may in fact degrade the estimate.

This study selected three closest control stations to estimate the missing data at the target station. An additional reason to select only three control stations is due to the availability of data. Since the dataset contains a few real missing data, the data records that have missing data must be removed. The number of available data records decreases when the number of control stations increases. Thus, the use of three control stations is deemed to be an appropriate selection for this study. However, it does not necessarily mean it is the best.

The rainfall data range from 1981 to 2001. The data from 1981 to 1998 are used to calibrate the models, and data from 1999 to 2001 are used to validate the developed models. Since there are a few real missing data records in control stations in the earlier period, such records have been removed. After removing missing records from calibration data, the proportion between validation and calibration data falls between 18 to 20 percents approximately. To validate the models, Mean Absolute Error (MAE) is adopted as given in equation (1).

$$MAE = \sum_{i=1}^{m} |Oi - Pi| / m \,. \qquad (\mathbf{1})$$

where O_i and P_i is the observed the estimated value respectively, m is the number of missing data.

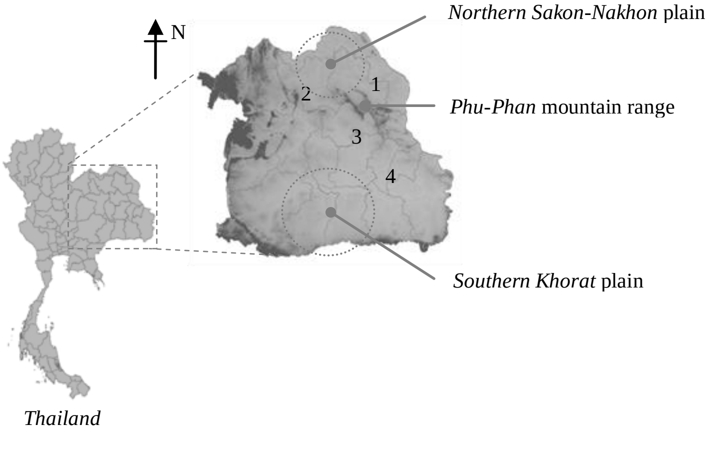

Fig. 1. Four selected case study sites in the northeast region of Thailand, case 1: ST356010, case 2: ST381010, case 3: ST388002, case 4: ST407005. Case 1 and Case 3 sites are located over and under the *Phu-Phan* mountains range. Case 2 sites are located in the *Northern Sakon-Nakhon* plain and Case 4 sites are located in the *Southern Khorat* plain.

4 The Modular Artificial Neural Networks

Figure 2 shows an overview of the proposed model. The proposed methodology could be divided into two steps. The first step is to partition the data and create the estimation modules. In this step, the training data are clustered into different groups and then the data in each group are used to train an ANN. The second step is to create an aggregation module. The function of this module is to finalize the decision value from those networks. In this study two aggregation methods are introduced. Both methods are based on the concept of Tobler's first law, "*Everything is related to everything else, but near things are more related than distant things*" [8].

In the first step, **s**ince the nature of rainfall data could be divided into dry and wet period, the proposed method partitions the data into two clusters according to the seasons. All the input-output pairs are then clustered by using Fuzzy C-Mean (FCM) clustering technique. Once the two training data are prepared, supervised neural network are used to capture the relationship between these input-output pairs. Among several types of supervised neural network, Back-Propagation Neural Network (BPNN) has been widely used in hydrological study. In this study one hidden layer BPNN is used to learn from the training data. The numbers of input node, hidden node and output node are three, four and one respectively. The transfer functional used in the model is sigmoid function.

The second step is to create an aggregation module. This study proposed two aggregation methods, *Inverse Distance Weighting Method* (MANN-IDWM) and *Optimized Weight of Subspace Reconstruction Method* (MANN-OWSR). In the first method, the final decision output should be closer to the decision output from closer ANN than farther ANN. The distance between the data point and the center of clusters are used to weight the final decision value from both ANNs inversely. The mathematic formula of MANN-IDWM is

$$z_o = [z_1 \frac{1}{d_1^k} + z_2 \frac{1}{d_2^k}] / \left[\frac{1}{d_1^k} + \frac{1}{d_2^k}\right] . \qquad \textbf{(2)}$$

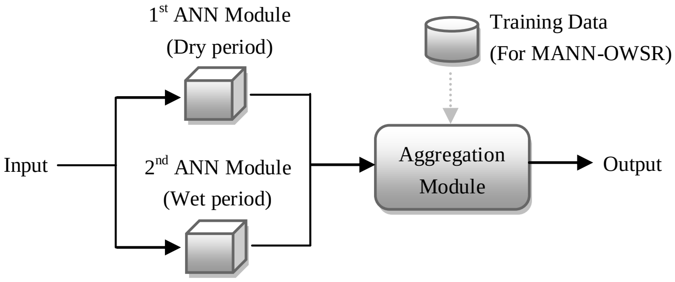

Fig. 2. The architectural overview of the proposed model. The first and second ANN captures the relationship of rainfall in the dry and wet period respectively. In the aggregation module, the training data are used only for the MANN-OWSR model.

where Z_1 is predicted value from ANN$_1$ and Z_2 is the predicted value from ANN$_2$, d_1 and d_2 are the distance between the data point to the centroid of cluster 1 and cluster 2 respectively. k is the power parameter. The optimized k parameter can be found from training data.

In the second method, the optimized weight of subspace reconstruction method, based on the idea that if the weight assigned to a data point in order to weight final decision value from both ANNs is optimal, this weight value should also be optimal for the nearest data points in the same manner. Assume δ to be a small region around an input vector Z_s and a set of data points $\{Z_1, Z_2,\ldots,Z_k\}$ to be the data points around the region δ in which

$$\|Z_i - Z_s\| \ll \delta\ . \tag{3}$$

If the weight applied to Z_s is the optimal weight. That weight should be the optimal value for all the points in the region. So, if the weight applies to all the points in that region is optimal, the error of equation shown below should be minimal.

$$\varepsilon = \frac{1}{k}\sum_{i=1}^{k}(z_i^{'} - z_i)^2\ . \tag{4}$$

where ε is mean square error, k = number of data point in the region δ, $z^{'}$ is predicted value from MANN and z is the observed value. Considering this case study, the final decision value comes from two ANNs. The final estimated value is $z^{'} = \alpha z_d^{'} + \beta z_w^{'}$ and $\alpha + \beta = 1$ Then

$$z^{'} = \alpha z_d^{'} + (1-\alpha)z_w^{'}\ . \tag{5}$$

where $z^{'}$ is final predicted results and α is weight applied. Replace equation (5) into equation (4). Then

$$\varepsilon = \frac{1}{k}\sum_{i=1}^{k}((\alpha z_{di}^{'} + (1-\alpha)z_{wi}^{'}) - z_i)^2\ . \tag{6}$$

The equation (6) is the cost function that we have to minimize in order to find the optimal value of α. Then the problem is to optimize one variable equation. In order to optimize the cost function, this study uses a MATLAB function call "*fminbnd*" to minimize MSE. The function "*fminbnd*" is used to find the minimum of the single variable function of a fix interval. It finds a minimum for a problem specified by $\min_x f(x)$, subject to $x_1 < x < x_2$ where x_1, x, x_2 are scalars and $f(x)$ is a function that returns a scalar. Its algorithm is based on golden section search and parabolic interpolation. More details have been described in references [9] and [10].

In the case that the data points in the region are sparse or there is no point in the defined region, the aggregation method will use MANN-IDWM instead. Another consideration is the size of the small region (or radius). This optimal size can be found by direct search using training data. However, for the rainfall data, the distribution of rainfall among a year is varying, so the radius should not be fixed in the input space.

Taking the distribution of data in Figure 3 into account, one can see that the distribution of data is concentrated near origin and spread out in all dimension. Thus, the proposed method partition input space into three regions. Region 1 begins from the origin to center of the first cluster. Region 2 is between center of first and second cluster. Region 3 is the area outside center of cluster 2. In each region, the training data have been used to investigate the appropriate radius by direct searching.

5 Experimental Results

To evaluate the accuracy of the developed models, the rainfall data from 1999 to 2001 are assumed to be missing data records and they needed to be estimated. The proposed models have been compared with the Inverse Distance Weighting Method (IDWM), the Correlation Coefficient Weighting Method (CCWM) and Artificial Neural Network (ANN). Table 1 shows the results of evaluation.

In IDWM, the optimized power parameter *k* could be defined by considering MAE of data in the calibration period when increasing power parameter. It was found that the optimized power parameters are 0.8, 4.5, 2.8 and 0 for case 1 to case 4 respectively. In CCWM, the correlation coefficient of rainfall data between each control stations and target station in calibration period are used in this method. The network architecture of the ANN is the same as the architecture used in the MANN method.

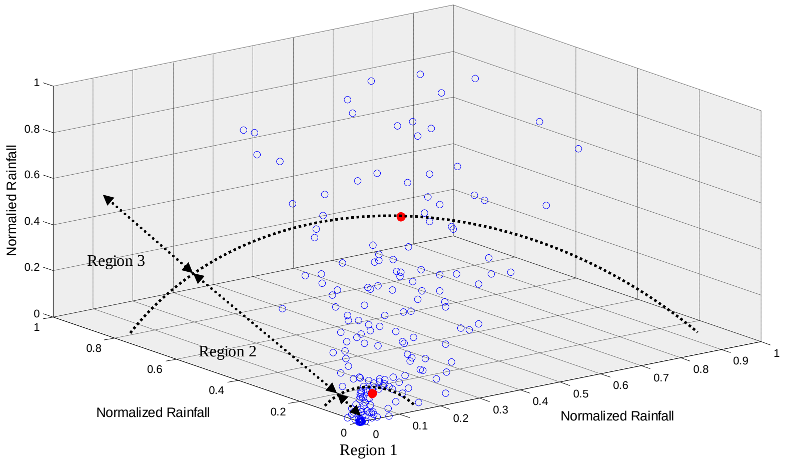

Fig. 3. An example of the distribution of rainfall data in the input space (TS356010).

Table 1. Mean Absolute Error (MAE) of validation data

Models	ST356010	ST381010	ST388002	ST407005
IDWM	245.02	267.32	500.46	399.10
CCWM	261.05	285.38	484.92	399.02
ANN	244.59	258.13	487.95	481.11
MANN - IDWM	232.04	230.83	456.25	387.52
MANN - OWSR	212.91	228.40	448.36	389.87

In case 1 (ST356010), CCWM gave the highest estimation error. IDWM and ANN showed no different in the accuracy. MANN-IDWM provided an improvement from ANN and other conventional method up to 5 percents. In turn, MANN-OWSR provided significantly improvement from MANN-IDWM to almost 8 percents. This case study pointed out that the proposed methods, especially MANN-OWSR can improve the performance of ANN and other conventional models.

In case 2 (ST381010), CCWM provided the lowest accuracy. IDWM provided better estimation than CCWM, and ANN provided better estimation than IDWM. MANN-OWSR showed a slight improvement over MANN-IDWM. However, both models showed better result than CCWM, IDWM and ANN of up to 13 percents approximately.

In case 3 (ST388002), IDWM provided the lowest accuracy whereas CCWM and ANN provided almost similar performance. MANN-IDWM improved the estimation of the ANN by 6.50 percents and MANN-OWSR improved the estimation of ANN by 8 percents. In this case study, MANN-OWSR again showed better estimation results than MANN-IDWM

In case 4 (ST407005), IDWM and CCWM provided almost similar estimation results whereas ANN showed very high estimation error in this case study. However, both MANN-IDWM and MANN-OWSR still provided lower estimation error than IDWM and CCWM. This case study pointed out that MANN-IDWM and MANN-OWSR could still perform good estimation results even though ANN provided high estimation error.

Since the large estimation error occurred to ANN in case 4, then, more investigation is needed. It was found that there are some rainfall records in the calibration period which the relationship of control stations and target station could be considered as irregular events; For example, there is an overshoot rainfall record at target station whereas rainfall data at control stations are normal. If such record occurred frequently in the training data, ANN could not provide reasonable estimation and thus yield large MAE. However, only ANN is affected by this noise data because ANN used this record as input-output pair in adapting process whereas the IDWM and CCWM do not use the output. In case of MANN, these irregular data are separated into two datasets. Although one ANN is affected by this data, another ANN is not. Therefore, when the final decision value is evaluated from both ANN, the irregular effect is reduced.

6 Conclusion

This study proposed the use of modular artificial neural networks to estimate the missing monthly precipitation records. The proposed models use fuzzy c-mean clustering technique to partition the data into dry and wet period according to the nature of the data. Back-propagation neural networks have been used to capture the relationship of rainfall in each period. In the aggregation module, this study used an inverse distance weighting method and an optimized weight of subspace reconstruction method to form the final decision value. Four case studies in the northeast region of Thailand have been used to test the proposed models. The simultaneous rainfall records from three nearest control stations were used to estimate the missing rainfall record at the target station. The experimental results reported so far have showed that the use of modular artificial neural network can improve the performance of single artificial neural network and other conventional method to estimate the missing rainfall data. Furthermore, modular artificial neural networks are more tolerant to irregular data than single artificial neural networks.

References

1. Kim, J., Pachepsky Y.A.: Reconstructing Missing Daily Precipitation Data using Regression Trees and Artificial Neural Networks for SWAT Streamflow Simulation. J. Hydrol. 394 (2010) 305–314
2. Teegavarapu, R.S.V., Chandramouli, V.: Improved Weighting Methods, Deterministic and Stochastic Data-driven Models for Estimation of Missing Precipitation Records. J. Hydrol. 312 (2005) 191–206
3. Teegavarapu, R.S.V.: Use of Universal Function Approximation in Variance-dependent Surface Interpolation Method: An Application in Hydrology. J. Hydrol. 332 (2007) 16–29
4. Teegavarapu, R.S.V., Tufail, M., Ormsbee, L.: Optimal Functional Forms for Estimation of Missing Precipitation Data. J. Hydrol. 374 (2009) 106–115
5. Piazza, A.D., Conti, F.L., Noto, L.V., Viola, F., Loggia, G.L.: Comparative Analysis of Different Techniques for Spatial Interpolation of Rainfall Data to Create a Serially Complete Monthly Time Series of Precipitation for Sicily, Italy. Int. J. Appl. Earth Obs. Geoinf. 13 (2011) 396–408
6. Kajornrit J., Wong, K.W., Fung, C.C.: Estimation of Missing Rainfall Data in Northeast Region of Thailand using Spatial Interpolation Methods. AJIIPS. 13(1) (2011) 21–30
7. Eischeid, J.K., Pasteris, P.A., Diaz, H.F., Plantico, M.S., Lott, N.J. Creating a Serially Complete, National Daily Time Series of Temperature and Precipitation for the Western United States. J. Appl. Meteorol. 39 (1999) 1580–1591
8. Miller, H.J.: Tobler's First Law and Spatial Analysis. A. Assoc. Am. Geog. 94(2) (2004) 284–289
9. Forsythe, G.E., Malcolm, M.A., Moler, C.B.: Computer Methods for Mathematical Computations. Prentice-Hall (1976)
10. Brent, R.P.: Algorithms for Minimization without Derivatives. Prentice-Hall, New Jersey (1973)