Hindawi Publishing Corporation
Advances in Artificial Intelligence
Volume 2014, Article ID 716398, 5 pages
http://dx.doi.org/10.1155/2014/716398

Hindawi

Research Article

Estimation of Missing Rainfall Data Using GEP: Case Study of Raja River, Alor Setar, Kedah

Nor Zaimah Che Ghani,[1] Zorkeflee Abu Hasan,[2] and Lau Tze Liang[3]

[1] *River Engineering and Urban Drainage Research Centre (REDAC), Universiti Sains Malaysia, Engineering Campus, Seri Ampangan, 14300 Nibong Tebal, Penang, Malaysia*
[2] *REDAC, Universiti Sains Malaysia, Engineering Campus, Seri Ampangan, 14300 Nibong Tebal, Penang, Malaysia*
[3] *School of Civil Engineering, Universiti Sains Malaysia, Engineering Campus, Seri Ampangan, 14300 Nibong Tebal, Penang, Malaysia*

Correspondence should be addressed to Nor Zaimah Che Ghani; zaimahcg@gmail.com

Received 15 May 2014; Accepted 1 August 2014; Published 9 September 2014

Academic Editor: Adel M. Alimi

Water resources and urban flood management require hydrologic and hydraulic modeling. However, incomplete precipitation data is often the issue during hydrological modeling exercise. In this study, gene expression programming (GEP) was utilised to correlate monthly precipitation data from a principal station with its neighbouring station located in Alor Setar, Kedah, Malaysia. GEP is an extension to genetic programming (GP), and can provide simple and efficient solution. The study illustrates the applications of GEP to determine the most suitable rainfall station to replace the principal rainfall station (station 6103047). This is to ensure that a reliable rainfall station can be made if the principal station malfunctioned. These were done by comparing principal station data with each individual neighbouring station. Result of the analysis reveals that the station 38 is the most compatible to the principal station where the value of R^2 is 0.886.

1. Introduction

The importance of precipitation is (1) identifying precipitation characteristics, occurrence and temporal and spatial variability, (2) statistical modeling and forecasting of precipitation, and (3) resolving the problems such as floods, droughts, and landslides as stated by Silva et al. [1]. But, in some cases, a large number of stations could be down at the same time, thus creating many inaccurate readings or missing data [2, 3]. In Malaysia, the number of rain gauge stations with complete records for a long duration is very scarce. Rainfall records often contain missing data values due to malfunctioning of equipment and severe environmental conditions. Thus, the estimation of rainfall is needed if missing data happened at the principal rainfall station. This study was to investigate the possibilities of correlating monthly rainfall of principal rainfall station to its six neighbouring stations. This was done to ensure that a reliable rainfall station can be done before proceeding with water resources management and flood management modelling.

1.1. Description of the Study Area. The study area was carried out in Alor Setar city, the capital of Kedah state in Malaysia. It is located within the Raja River catchment. It is prone to flood due to its flat and low elevation. In 1992, Department of Irrigation and Drainage (DID) carried out the Flood Mitigation Project to solve the flooding problems of Alor Setar city where the whole Raja River system was converted to concrete lined channel. It was separated from Kedah River by gated structure and pumping station [4].

A study is being conducted to investigate how Raja River system responds to the land use change by carrying out hydrologic and hydraulic modeling. One of the main inputs of the modeling is precipitation data but missing precipitation data has always been an issue for hydrologic modelling as stated earlier. There are seven rainfall stations in this study area; station 6103047 as principal station is surrounded by six Muda Agricultural Development Authority (MADA) rainfall stations as shown in Figure 1 and Table 1. The minimum densities recommended of precipitation stations by the WMO are 1 station for 250 km^2 for the mountainous

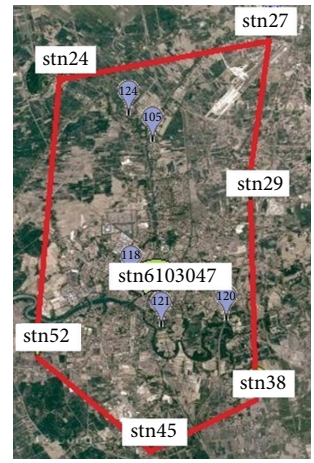

Figure 1: Location map of a study area. Close-up view of the study area in Alor Setar, Kedah.

area, 1 for 900 km^2 for the coastal area, and 1 for 10 km^2 for urban areas [5]. In the study area, there are seven stations within 200 km^2 for the study area (approximately 1 station for 30 km^2).

2. Data and Methodology

In order for hydrologic modelling to be conducted smoothly, data consistency of a principal station was compared to its neighbouring rainfall station by applying gene expression programming (GEP) technique. Monthly rainfall series data have been obtained from DID and MADA for 9-year periods from 2001 to 2009. For this study, MADA stations were selected based on closest distance with station 6103047 as shown in Table 1.

Genetic programming (GP), a branch of genetic algorithms (GA), is a method for determining the most "fit" computer program by artificial evolution. GP initializes a population which consists of chromosomes, and the fitness of each chromosome is evaluated regarding a target value. The individuals in the new generation are, in their turn, through a few developmental processes, such as expression of the genomes, confrontation of the selection environment, and reproduction with modification. The reproduction includes not only replication but also the action of genetic operators capable of creating genetic diversity. During replication, the genome is copied and transmitted to the next generation. So, in GEP, a chromosome might be modified by one or several operators at a time or might not be modified at all [6–10].

Hashmi et al. [11] show simple example of a GEP model having two genes (terms), which are linked by an addition function, and presented here to clarify the working of the GEP system. This GEP chromosome is given by

$$(a * b) + \left(\frac{c}{d}\right), \tag{1}$$

where "a," "b," "c," and "d" are predictor variables and +, $*$, and / represent addition, multiplication, and division, respectively. Equation (1) can also be expressed by the following expression tree (ET) which is usually produced by GEP software packages. In this study, the data for the training set in GEP is selected from 2001 to 2006 and the rest is used as the testing set. The functional set and operational parameters used in the present GEP modelling are listed in Tables 2 and 3, respectively.

3. Results and Discussion

GEP was used to predict precipitation of station 6103047 using 9 years of monthly rainfall data to select the most suitable rainfall station. By using GEP, (2) was generated where x refers to station 38. The equation for GEP also can be expressed by expression tree (ET) as shown in Figure 2 to show the relationship between station 38 and station 6103047. Only station 38 is discussed here since it is the best rainfall station. Consider

$$\begin{aligned}
&\text{stn } 6103047 \\
&\quad = \tan\left[\left(\sin x - x^2\right)\left(x^2 - 0.840185\right)\right] \\
&\qquad + (\ln)(\exp)(\sin)\left[\sqrt{x} - \sqrt{(x + 8.399719) - x}\right] \\
&\qquad + \left[(x + 0.154083) - (-2.685944)^2\left(\exp x^2\right)\right] \\
&\qquad + \sin(\cos -5.893769) + (10.101483 \tan x)^2 \\
&\qquad \times \left[\exp(-232.026432 + \cos x)\right] \\
&\qquad + \exp\left(\left(\left(\cos(x - 0.103241)^3\right)^2\right)^3\right) \\
&\qquad + \tan\left[\frac{\sin x}{6.750244} - (\cos(9.078826x))\right]^2.
\end{aligned} \tag{2}$$

The GEP was able to determine the most suitable rainfall station to replace the principal station. The coefficient of determination (R^2) and the root mean square error (RMSE) are used in the current study. The R^2 represents the degree of

Table 1: Details of the rainfall stations.

Station name		Custodian	Coordinate		Distance station 6103047-MADA station (km)
			Latitude	Longitude	
Stor JPS	Station 6103047	DID	6.105556 N	100.3917 E	—
Telok Chengai	Station 52	MADA	6.097806 N	100.3315 E	4.3
Bt. 3 Tandop	Station 45	MADA	6.068278 N	100.3676 E	5
Alor Penyengat	Station 38	MADA	6.085 N	100.401 E	5
Hutan Kampong	Station 29	MADA	6.149028 N	100.399 E	5.3
Kepala Batas	Station 27	MADA	6.201445 N	100.4047 E	10.6
Gunung Keriang	Station 24	MADA	6.188888 N	100.3388 E	8.8

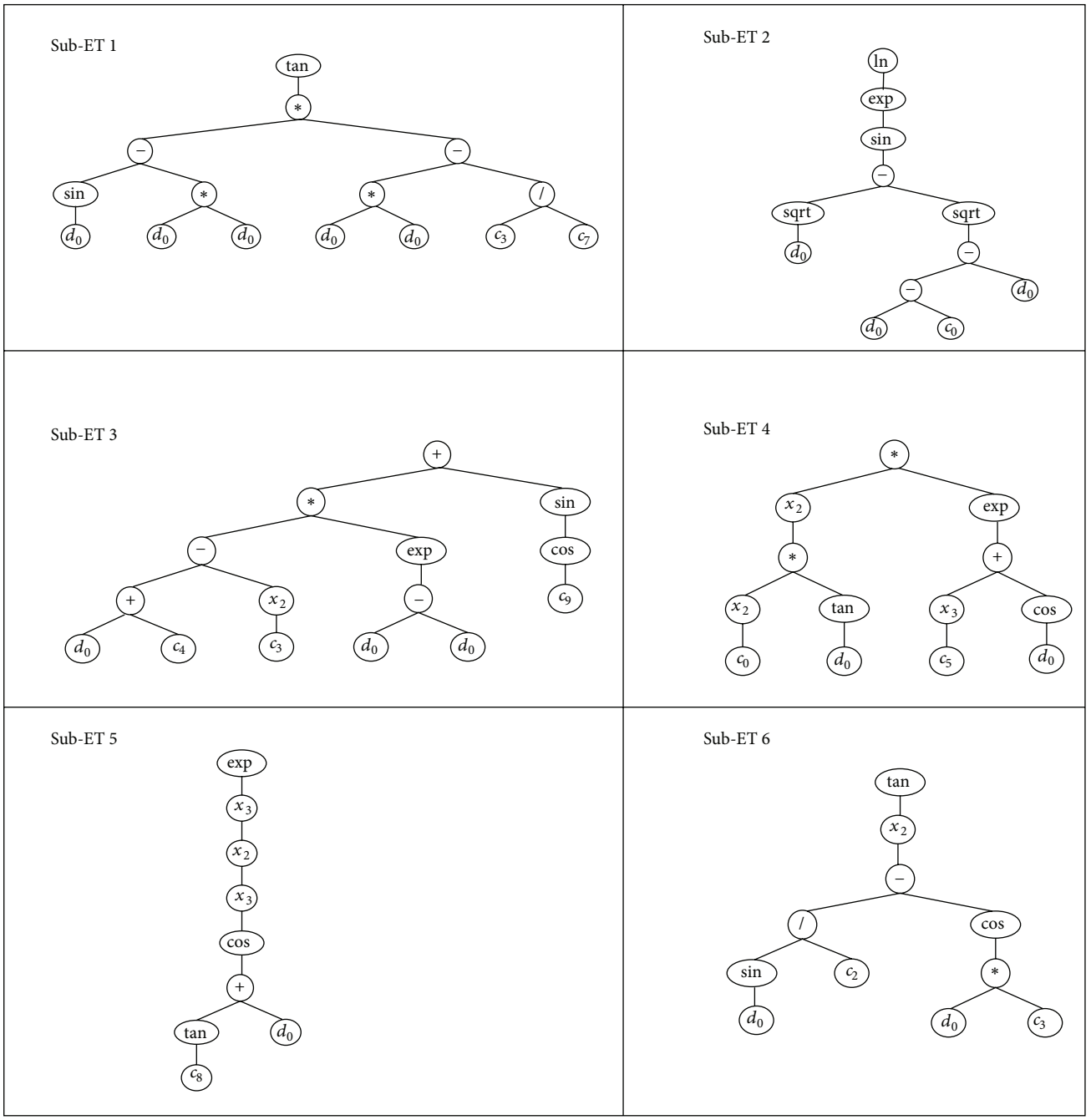

Figure 2: Expression tree shows the relationship between station 38 and station 6103047.

Table 2: Functional set for the GEP model.

Function Set	Symbol	Weight	Arity
Addition	+	2	2
Subtraction	−	2	2
Multiplication	*	2	2
Division	/	1	2
Square root	sqrt	1	1
Exponential	exp	1	1
Natural logarithm	ln	1	1
X to the power of 2	$X2$	1	1
X to the power of 3	$X3$	1	1
Sine	sin	1	1
Cosine	cos	1	1
Tangent	tan	1	1

Table 3: Genetic operators used in GEP modeling.

Parameters	Definition	Value
P_1	Mutation rate	0.044
P_2	Inversion rate	0.1
P_3	IS transposition rate	0.1
P_4	RIS transposition rate	0.1
P_5	One-point recombination rate	0.3
P_6	Two-point recombination rate	0.3
P_7	Gene recombination rate	0.1
P_8	Gene transposition rate	0.1

Table 4: Summary of the analysis from 2001–2009 by using GEP.

Station	R	R^2
Station 6103047-Station 52	0.895	0.802
Station 6103047-Station 45	0.931	0.867
Station 6103047-Station 38	0.941	0.886
Station 6103047-Station 29	0.926	0.857
Station 6103047-Station 27	0.864	0.747
Station 6103047-Station 24	0.840	0.706

association between the predicted and the measured values as shown in

$$R^2 = \left[\frac{\sum xy}{\sum x^2 \sum y^2}\right]^2. \quad (3)$$

The R^2 of GEP technique (0.886) for station 38 in Table 4 has the highest values. If R^2 close to 1, it indicates that we have accounted for almost all the variability with the variables specified in the model.

The values of R for GEP technique of station 38 is 0.941. Its value of 1 represents a perfect relation, and 0 indicates no relationship between the variables. The degree to which two or more predictors are related to the dependent variable is expressed in R. The function has a determination coefficient as a measure of the goodness of fit of the model, and this represents the proportion of the variation of the dependent variable (station 6103047 rainfall depth).

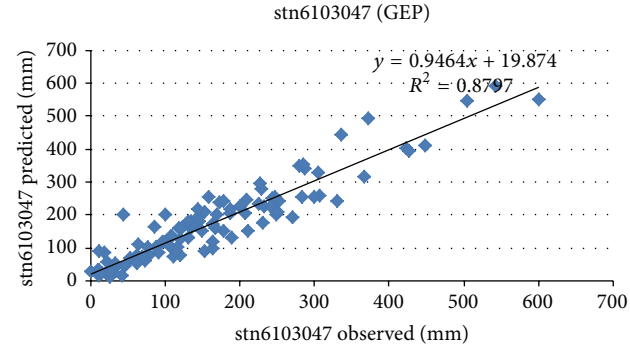

Figure 3: Observed and predicted graph for station 6103047 using GEP technique.

The graph in Figure 3 shows the predicted station 6103047 against the observed station 6103047 which achieves acceptable R^2. The value of R^2 is 0.879 for GEP technique. So, station 38 for GEP technique is reasonably close to the observed station 6103047 as the R value is 0.941. The larger R value, the stronger the association between the two variables and the more accurate the prediction of the values of station 6103047.

4. Conclusions

This study is using GEP technique to determine the most fitted rainfall station to the principal rainfall station. The predicted GEP model gives satisfactory results. As GEP technique provides more efficient result, it will be used to estimate the missing rainfall and to correlate monthly precipitation data from the principal station to station 38. From the analysis, station 38 is the most fitted rainfall to the principal station as having the highest R^2 (0.886) which is very close to 1, suggesting very little discrepancy between observed and predicted precipitation. It shows that GEP can be used as an effective tool to be used for estimating precipitation.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors are grateful to the Universiti Sains Malaysia for providing research Grant (1001/REDAC/814085) to study integrated river basin management for Raja River, Kedah, Malaysia.

References

[1] R. P. de Silva, N. D. K. Dayawansa, and M. D. Ratnasiri, "A comparison of methods used in estimating missing rainfall data," *Journal of Agricultural Science*, vol. 3, pp. 101–108, 2007.

[2] R. S. V. Teegavarapu, M. Tufail, and L. Ormsbee, "Optimal functional forms for estimation of missing precipitation data," *Journal of Hydrology*, vol. 374, no. 1-2, pp. 106–115, 2009.

[3] J. Kajornrit, K. W. Wong, and C. C. Fung, "Estimation of missing rainfall data in Northeast region of Thailand using Kringing methods: a comparison study," in *Proceedings of the International Workshop on Bio-Inspired Computing for Intelligent, Environments and Logistic Systems*, pp. 1–8, 2011.

[4] M. S. Ramli, Z. Abu Hasan, and K. Hock Lye, "Application of one-dimensional water quality modelling for in stream dissolved oxygen," in *Sustainable Solutions for Global Crisis of Flooding, Pollution and Water Scarcity*, pp. 1–149, 2011.

[5] WMO (World Meteorological Organisation), "Hydrology: from measurement to hydrological information," in *Guide to Hydrological Practices*, vol. 168, WMO, Geneva, Switzerland, 6th edition, 2008.

[6] C. Ferreira, "Gene expression programming: a new adaptive algorithm for solving problems," *Complex Systems*, vol. 13, no. 2, pp. 87–129, 2001.

[7] C. Ferreira, *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*, Springer, 2nd edition, 2006.

[8] H. M. Azamathulla and Z. Ahmad, "Gene-expression programming for transverse mixing coefficient," *Journal of Hydrology*, vol. 434-435, pp. 142–148, 2012.

[9] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, University of Michigan Press, 1975.

[10] H. Mohammad, A. Amin uddin, A. Ghani, C. Siang, L. Chun, and K. Chang, "Gene-expression programming for the development of a stage-discharge curve of the Pahang River," *Water Resources Management*, vol. 25, no. 11, pp. 2901–2916, 2011.

[11] M. Z. Hashmi, A. Y. Shamseldin, and B. W. Melville, "Statistical downscaling of watershed precipitation using Gene Expression Programming (GEP)," *Environmental Modelling and Software*, vol. 26, no. 12, pp. 1639–1646, 2011.