Chapter 61
Flood Frequency Analysis at Ungauged Sites in the Peninsular Malaysia Using Multiple Linear Regressions

Ani Shabri and Nur Shahidah Roselan

Introduction

Human society faces great problems due to extreme environmental events. For example, floods, rainstorms, droughts, and high winds that cause tornadoes and such destroy almost anything that is in their vicinity at the moment of occurrences. Flood, also known as deluge, is a natural disaster that could diminish properties, infrastructures, animals, plants, and even human lives.

In terms of the number of population affected, frequency, area extent, duration, and social economic damage, flooding is the most natural hazard in Malaysia. Malaysia has experienced major floods since 1920. These flood events occurred in various states in Peninsular Malaysia [1].

Flooding occurs when the volume of water exceeds the capacity of the catchment area. Floods are one of the natural disasters that occur not only in Malaysia but also in other parts of the world. It is also the most costly natural hazard due to its ability to destroy human properties and lives. The basic cause of river flooding is the incidence of heavy rainfalls, such as the monsoon season or convection, and the resultant large concentration of runoff, which exceeds river capacity [1].

The study of water-related characteristics and modeling throughout the Earth such as the movement, distribution, resources, hydrologic cycle, and quality of water is called hydrology. By knowing and analyzing statistical properties of hydrologic records and data like rainfall or river flow, hydrologists are able to estimate future hydrologic phenomena.

A. Shabri (✉) • N.S. Roselan
Mathematic Department, Faculty of Science, Universiti Teknologi Malaysia,
Johor Bahru 81310, Johor, Malaysia
e-mail: ani@utm.my; shieda_1405@yahoo.com

A. Kasim et al. (eds.), *Proceedings of the International Conference on Science, Technology and Social Sciences (ICSTSS) 2012*, DOI 10.1007/978-981-287-077-3_61,

Cluster analysis has become a common tool for the marketing researchers. Both the academic researcher and marketing application researchers rely on the technique for developing empirical groupings of persons, products, or occasions which may serve as the basis for further analysis. Despite its frequent use, little is known about the characteristics of available clustering methods or how clustering method should be employed. One indication of this general lack of understanding of clustering methodology is the failure of numerous authors in the marketing literature to specify what clustering method is being used [2].

Multiple Linear Regressions (MLR)

MLR is a method used to model the linear relationship between a dependent variable and one or more independent variables. The dependent variable is sometimes also called the predictand and the independent variables the predictors. MLR is based on least squares: the model is fit such that the sum of squares of differences of observed and predicted values is minimized.

MLR is probably the most widely used method in dendroclimatology for developing models to reconstruct climate variables from tree-ring series. Typically, a climatic variable is defined as the predictand and tree-ring variables from one or more sites are defined as predictors. The model is fit to a period – the calibration period – for which climatic and tree-ring data overlap. In the process of fitting, or estimating, the model, statistics are computed that summarize the accuracy of the regression model for the calibration period.

The performance of the model on the data not used to fit the model is usually checked in some way by a process called validation. Finally, tree-ring data from before the calibration period are substituted into the prediction equation to get a reconstruction of the predictand. The reconstruction is a prediction in the sense that the regression model is applied to generate estimates of the predictand variable outside the period used to fit the data. The uncertainty in the reconstruction is summarized by confidence intervals, which can be computed by various alternative ways.

The model expresses the value of a predictand variable as a linear function of one or more predictor variables and an error term (model 1):

$$y_i = b_0 + b_1 X_{i,1} + b_2 X_{i,2} + \cdots + b_k X_{i,k} + e_i \qquad (61.1)$$

The model (61.1) is estimated by least squares, which yields parameter estimates such that the sum of squares of errors is minimized. The resulting prediction equation is (model 61.2)

$$\hat{y}_t = \hat{b}_0 + \hat{b}_1 x_{t,1} + \hat{b}_2 x_{t,2} + \cdots + \hat{b}_k x_{t,k} \qquad (61.2)$$

Hierarchical Cluster Analysis (HCA)

Hierarchical clustering is a general approach to cluster analysis in which the object is to group together objects or records that are close to one another. A key component of the analysis is repeated calculation of distance measures between objects and between clusters once objects begin to be grouped into clusters.

The objective of cluster analysis is to assign observations to groups or clusters so that observations within each group are similar to one another with respect to variables or attributes of interest, and the groups themselves stand apart from one another [3].

The outcome is represented graphically as a dendogram. The initial data for the hierarchical cluster analysis of N objects is a set of $N \times (N-1)/2$ object-to-object distances and a linkage function for computation of the cluster-to-cluster distances.

The two main categories of methods for hierarchical cluster analysis are divisive methods and agglomerative methods. In practice, the agglomerative methods are of wider use. On each step, the pair of clusters with the smallest cluster-to-cluster distance is fused into a single cluster. The most common algorithm for hierarchical clustering is average linkage clustering.

Average Linkage Clustering

The average linkage clustering is a method of calculating distance between clusters in hierarchical cluster analysis. The linkage function specifying the distance between two clusters is computed as the average distance between objects from the first cluster and objects from the second cluster. The averaging is performed over all pairs (*x, y*) of objects, where *x* is an object from the first cluster and *y* is an object from the second cluster. Mathematically the linkage function – the distance between clusters *X* and *Y* – is described by the following expression (Model 61.3):

$$D\left(X,Y\right)=\frac{1}{N_x \times N_y}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} d\left(x_{i,}y_j\right); \tag{61.3}$$

Data Source

Peninsular Malaysia covers 131,000 km^2 or 39.7 % of the total land of the country. Malaysia's climate is categorized as equatorial, being hot and humid throughout the year. Rainfall occurs throughout the year with the total annual rainfall being within 2,000–4,000 mm. Malaysia faces two monsoon winds seasons – the Southwest Monsoon from late May to September and the Northeast Monsoon from November to March. The Northeast Monsoon brings in more rainfall compared to the Southwest [1].

This research utilizes the data of daily stream flow in Peninsular Malaysia, which are collected from the Department of Irrigation and Drainage, Ministry of Natural Resources and Environment, Malaysia. The analysis of data focuses on estimating the annual maximum flow series that measure the peak of flow discharge for each year. The data that contains *n* observed years per site has a sample size of *n*. Basic information of each 70 catchments are the catchment's area (CA), elevation (E), longest drainage path (DP), mean rainfall (MR), and river slope (RS). 56 data are used as training data and another 14 data are used for testing data.

Results and Discussion

MLR Models

The MLR models that are considered in this paper are

$$\begin{aligned}
\hat{y}_1 &= a_0 + a_1 CA + a_2 DP + a_3 MR + a_4 E + a_5 RS + \varepsilon \\
\hat{y}_2 &= a_0 + a_1 CA + a_2 DP + a_3 MR + a_4 E + \varepsilon \\
\hat{y}_3 &= a_0 + a_1 CA + a_2 DP + a_3 MR + \varepsilon \\
\hat{y}_4 &= a_0 + a_1 CA + a_2 DP + \varepsilon \\
\hat{y}_5 &= a_0 + a_1 CA + \varepsilon \\
&\vdots
\end{aligned}$$

and the best model for this study is

$$\hat{y} = a_0 + a_1 CA + a_2 DP + a_3 MR + \varepsilon \text{ which is}$$

$$\hat{y} = -21.5861 + 0.075209 CA + 4.290698 DP + (-2.68401) MR$$

To analyze these models further, the statistical measurements of the MLR and MLR with HCA are compared. The performances of all the models are in Table 61.1.

The objective of this paper is to assess the performance of the MLR model with HCA in estimating the flood quantiles for ungauged sites in Peninsular Malaysia. The models are developed for 10-, 50-, and 100-year quantiles. The three flood quantiles and the two models used for comparison purposes are applied to the study case database. For MLR and MLR with HCA models, the results obtained is presented in Table 61.1.

The MSE and RMSE of an estimator are the expected value of the square of the error. The error is the amount by which the estimator differs from the quantity to be estimated. The smaller the mean squared error is, the closer the estimator is to the actual data. Small mean squared error means that the randomness reflects the data more accurately than a larger mean squared error. Based on Table 61.1, it shows that MLR with HCA model is performed better compared to the MLR model.

Table 61.1 Forecasting performance indices of models for MLR and MLR with HCA

Quantile	$T = 10$ years		$T = 50$ years		$T = 100$ years	
Models	MLR	MLR with HCA	MLR	MLR with HCA	MLR	MLR with HCA
MSE	3.158313	**1.445402**	3.376043	**1.704397**	3.393319	**1.328771**
MAE	1.541323	**0.941796**	1.588502	**1.075966**	1.560718	**1.001687**
MAPE	4.023795	**4.773954**	4.255211	**5.263682**	4.558945	**5.513084**
RMSE	1.777164	**1.054645**	1.837401	**1.258243**	1.842096	**1.119105**

The MAE is the average over the verification sample of the absolute values of the differences between forecasts and the corresponding observations. The MAE is the linear score which means that all the individual differences are weighted equally in the average. The smaller MAE means that the forecasts' value is closer to the observed value compared to larger MAE. In Table 61.1, the results shows that the MLR with HCA model is performed better compared to MLR.

The MAPE is the measure of accuracy of a method for constructing fitted time series values in statistics, specifically in trend estimation. It usually expresses accuracy as a percentage. The smaller value of MAPE shows that the model is performed better compared to the other model which means the MLR with HCA model is better than MLR model.

Conclusions

To illustrate the capability of the MLR with HCA model, this model is compared to the MLR model. For modeling study, hydrologic and physiographic data from using 70 catchments in the Peninsular Malaysia were used. The flood quantile associated with 10-, 50-, and 100-year return periods was considered. The overall performance of each model is examined using MSE, RMSE, MAE, and MAPE. The comparison between the two models shows that the MLR with HCA model performance is better.

Instead of these variables for hydrological modeling studies, it is of course extremely probable that the predictions would present more accuracy provided that more available other variables such as physiography, soil and land use properties, and climate were included in the modeling setup. The further studies for the same region are recommended including more available independent variables in the modeling setup.

References

1. Ministry of Natural Resources and Environment, June 2007.
2. Girish, P., & David, W. (1993). *Journal of Marketing Research*, *XX*.

3. Peter, T. (1997). *Cluster analysis*.
4. Paul, E. R., & William, M. F. (2004, June). *Applications of a MLR model to the analysis of relationships between eastward- and westward-moving intraseasonal modes*.
5. Attua, E. M. (2008). Using multiple linear R techniques to quantity carbon stocks of fallow vegetation in the tropics. *West African Journal of Allplied Ecology, 12*, 1–7.