ANALYSIS OF FLOOD FREQUENCY USING ANN

[1]Frederik DARA, [2]Skënder OSMANI

[1]Department of Mathematics, Faculty of Natural Sciences, University of Tirana, Albania
frederik.dara@fshn.edu.al
[2]Department of Energy Resources, Faculty of Geology and Mining, Polytechnic University, Tirana, Albania
s_osmani@yahoo.com

Abstract

The one of the problems that continues to be faced by the practicing Engineering Hydrologist is the estimation of design flood quantiles for ungauged locations. Many studies has been adopted the methodology based on the method of residuals. This method has been enough criticized for its dependence on geographical vicinity. Multivariable techniques, such as clustering analysis have weakness also. Recently, an increasely attention has been attended to the Artificial Neural Networks, both for classifying basins based upon their mapped characteristics, and for relating flood magnitudes or parameters of the at-sites frequency distribution of floods to those basin characteristics. Applications of Artificial Neural Networks to flood data has demonstrated that clusters of locations can be detected that were not visually obvious from mapping the residuals between observed index floods and those computed from a regional regression equation. In this paper, these two approaches are applied to flood data and compared and compared with the results obtained from applying the traditional MLRA-based methodology. The traditional methods was result in an equation that explain 84.6 % of the variance of MAF. The use of a multilayer perceptron was found to be superior to a regression equation due to reducing the root mean square error of index flood magnitudes for locations that are not included in Neural Network training.

Keywords: Artificial Neural Network, multilayer perceptron, regression, flood, residuals.

Introduction

The day-to-day work of Hydrologist is heavily depend upon the manipulation and interpretation of recorded data. In general, the longer the period of record, the lower the standard errors of estimate of hydrological design variables. The surge in hydrological research activities that occurred between 1930 and 1955 was attributed by Linsley (1967) to the wider availability of records from the systematic measuring programmes begun earlier in the 20th century. A further stimulus to the development of hydrometric networks occurred during the International Hydrological Decade from 1865-1974. Unfortunately, the last 25 years have seen a decline in network densities in many parts of the world, owing to civil unrest or lack of trained personnel and resources for operation and maintenance. The degree of deterioration has been such that the World Bank (1993) has warned that inadequate and unreliable data constitute a serious constrain to developing country-wide water resources strategies and managing water efficiently.

Hydrologists have responded to this situation by developing increasingly sophisticated methods for the regionalization of hydrological variables. The method of residuals, as applied to regional flood frequency analysis (Dalrymple, 1960), provides a

seminal example of an approach that is still widely adopted (e.g. Benson, 1962; Gunter, 1974: Natural Environment Research Council, 1975; Nouh, 1988). The basic methodology addresses two problems:

1. how can hydrologically homogeneous regions be identified? and once identified
2. how might the dependent variables, such as the magnitude of an index flood, be related to independent variables that can be measured from a topographic map or drawn from national rainfall statistics?

In the method of residuals, problem (2) is approached by the use of multiple linear regression analysis (MLRA), applied initially to the combined data set of the region. The residuals, i.e. the differences between the computed and observed values of the index flood, are then mapped, and examined for clusters of sites with values t hat are similar in both mangitude and and sign. Such clusters are then designated hydrologically homogeneous sub-regions, and new equations for the index flood are derived for the reduced data sets.

This approach is obviously heavily dependent upon geographical proximity in defining the sub-regions. Some work has moved away from this constraint, with multivariate statistical methods, such as cluster analysis, being used to define classes of catchment and discriminant analysis being applied to allocate ungauged catchments to one of the identified classes. The drawbacks to such approaches have been well described by Nathan and McMahon (1990) and will not be reiterated here. However, the notation that a group does not have to be geographically close to form a sub-region (e.g. Wiltshire, 1985) or that a given site may have an affinity with a different set of sites for quantile estimation (e.g. Burn, 1990) has clearly been attracting increasing support. Some other studies (Hall *et al.*, 1998; Hall *et al*., 1999) have shown that hydrologically plausible results can be obtained by applying modern mathematical tools, such as Artificial Neural Networks (ANNs), to regional flood frequency analysis. In particular, catchments may be classified on the basis of their mapped characteristics by the use of unsupervised learning with a Kohonen self-organizing feature (SOM) map, and the parameters of the at-site frequency distributions of annual floods may be related to catchment characteristics through supervised learning with a multi-layer perceptron (MLP) type of ANN.

In this paper, these two approaches are applied to flood data from the islands of Java and Sumatra in Indonesia, and compared with the results obtained from applying the traditional MLRA-based methodology. The data set followed by a summary of the results obtained from the method of residuals. The ANN modeling is then outlined in followed section, with demonstrates the advantages to be gained both from the use of ANNs for estimating the parameters of the at-site frequency distributions of floods and from breaking the data set into clusters. The paper closes with some concluding remarks on possible refinement to the suggested methodology.

MATERIAL AND METHOD

Firstly, let us take a look at data set.

Data set

The data upon which this study was based were obtained from the Flood Design Manual for Java and Sumatra (DFMJS, 1983). The available records of floods in the Island of Java were examined during the first phase of the project, with a similar study of Sumatran data forming the second phase. The Data Appendix to the Manual provides information on the floods recorded at 50 sites in Java and 83 in Sumatra, along with a table of 11 catchment characteristics for each site. These data sets represent a situation typical of that facing a Hydrologist working in a developing country, with the majority of stations being

commissioned in the 1960s and 1970s and with long-term records few and far between. Of these data, 48 sites in Java and 44 in Sumatra were chosen for annual flood analysis (i.e. using the maximum instantaneous discharge in each water year), and one site in Java in 15 in Sumatra, for which partial duration series were available, were employed for independent testing purposes.

The method of residuals

The FD MJS proposed a 4-variable regression equation relating the mean annual flood, MAF (m^3/s), to catchment area, AREA (km^2), mean annual maximum catchment one-day rainfall, APBAR (mm), simple river slope, SIMS (m/km), and a dimensionless lake index, defined as the proportion of the area draining through a lake or storage, LAKE. Using the reduced data set off 92 stations, an alternative equation was derived using stepwise linear regression as follows:

$$MAF = 0.00013 \times AREA^{0.78} \times AAR^{1.241} \times (1 + PLTN)^{-1.79} \times (1 + LAKE)^{-2.282} \quad (1)$$

In this equation, AREA and LAKE are as defined above. AAR is the average annual rainfall (mm) for the catchment, and PLTN is a plantation index, defined as the proportion of AREA given over to plantations. Equation (1) explains 84.6 % of the variance of MAF. However, when the residuals for individual stations were mapped, no consistent spatial groupings in terms of their magnitudes and sings were discernible.

ANN modelling

In developing a general relationship between MAF and catchment characteristics, an MLP type of ANN can be developed in place of equation (1). In addition to the characteristics referred to above, values were available for main stream length, MSL (km), main stream slope, S1085 (m/km), a forest index, defined as the proportion of AREA covered by forest, FOREST, a paddy index, defined as the proportion of AREA covered by paddy rice, PADDY, and a swamp index, defined as the proportion of AREA covered by swamp, SWAMP. A twelfth variable, catchment shape, SHAPE, defined as quotient of AREA and the square of MSL, was added to the data set. The *NeuralSolutions* software was then applied to develop a series of MLP (three-layer perceptron) type of ANNs using the above-mentioned catchment characteristics as inputs and either the parameters of the at-site Extreme Value type I (EVI) distribution or the first and second probability weighted moments (PWMs) as outputs. In each case, after rejecting one site that was an obvious outlier, the networks ware trained on 66 sites and verified on 25 sites. A hyperbolic tangent activation function was employed, and the numbers of neurons in the hidden layer and the numbers of training epochs were varied in order to identify the best network architectures.

Separate ANNs were developed for different selections of catchment characteristics as input variables, ranging from all 12 down to 4. Ten ANNs were trained for each configuration. Figure 1 shows the variation of the average root mean square error (RMSE) based upon the verification data with the number of parameters. In order to investigate the consistency of the results over all ten trained networks, Figure 1 also shows a plot of confidence limits located at plus and minus one standard deviation about the average RMSE.

The ANN with the lower average RMSE of 253 m^3/s was obtained with 8 input variable. This figure can be combined with the 275 m^3/s produced by applying Equation (1) to its set of verification data (shows as a diamond on Figure 1). This indicates a general improvement by the use of the ANN. Indeed all of the ANNs produced RMSEs that were as good as or smaller than that of Equation (1). Comparable results, and so only results relating to outputs of the location and scale parameters of the EVI distributions are reported subsequently.

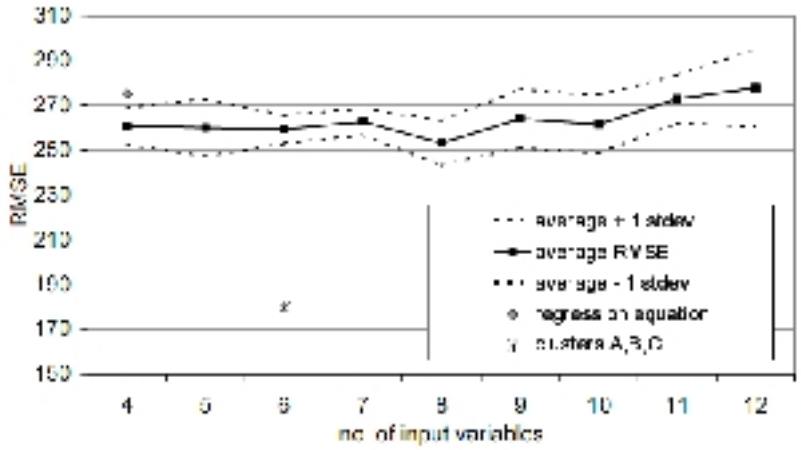

Figure 1. RMSEs for mean annual floods computed from a regional regression equation (MLRA) and various combinations of ANNs.

As a further test of the efficacy of the ANN models, estimates of MAF were computed for the 16 sites for which partial duration series estimates of the mean annual flood were available. For these sites, Equation (1) yielded a RMSE of 615 m^3/s, whereas the 8-variable ANN produced an RMSE of 373 m^3/s, an improvement of some 40 percent.

Having demonstrated a clear advantage for the use of an ANN to relate catchment characteristics and parameters of the at-site flood frequency distribution, the question arise as to whether further improvement is possible by clustering the catchments prior to developing the ANNs. For this purpose, the data were analyzed using a linear Kohonen SOM (Kohonen, 1995). In view of the relatively small data set available, the original 12 variables were reduced to six. S1085 was preferred to SIMS, and both AREA and MSL were included instead of SHAPE. FOREST, PADDY and SWAMP were excluded because of the low proportions of variance of the MAFs with which key were associated in the regression analysis, but PLTN and LAKE were included along with AAR but not APBAR. Since at least two, and possibly three, clusters were expected, 15 output neurons were included. Separate training runs were performed with neighbourhood radii ranging from 3 to 8, each training run being repeated several times to check for consistent results. Euclidean distance was used as the similarity measure.

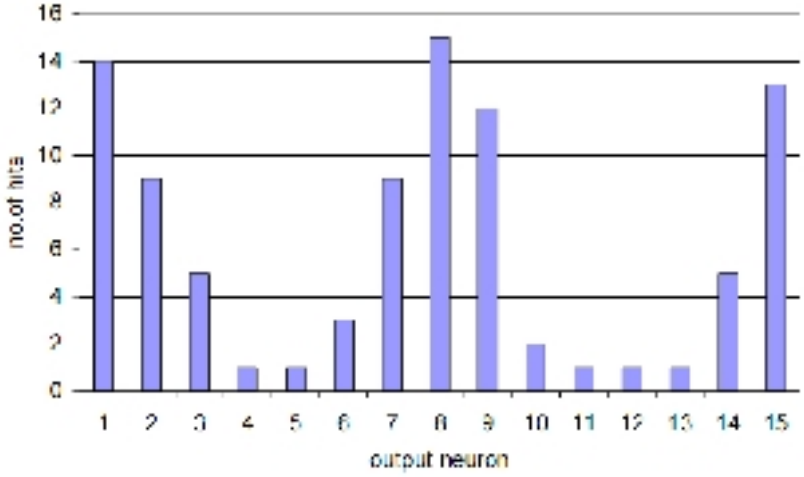

Figure 2: Count map of linear SOM map for six catchment characteristics

The results obtained showed that with small neighbourhoods radii, four or even five clusters were identifiable, but not all catchments were clearly classified and some clusters

were small and not consistently reproducible. In general, increasing the neighbourhood radius resulted in fewer clusters and smaller numbers of unclassified catchments. The best and most consistent results were obtained with a neighbourhood radius that started at 8 and which reduced to 5 during training. This yielded three well-defined, stable clusters with all catchments classified as shown in Figure 2.

The three clusters contained 29, 43 and 20 catchments respectively, each containing representatives from both Java and Sumatra. The weights associated with the six connections to each output node are the standardized cluster centers in Euclidean space, the de-standardised values of which may be regarded as the characteristics of Representative Regional Catchments (RRCs) defining each class. The weighted average of the weight values for each cluster are summarized in Table 1, which shows that the values of the separate characteristics change monotonically from class to class.

Table 1. Characteristics of the RRCs for the three clusters identified by a Kohonen network

Characteristic	Cluster A	Cluster B	Cluster C
AREA	388.9	861.5	3689.2
AAR	3291	2637	2509
MSL	41.95	66.37	144.08
S1085	35.95	14.89	5.97
1+PLTN	1.010	1.075	1.128
1+LAKE	1.001	1.005	1.080

Class A represents relatively small drainage areas with a high main channel slope and average annual rainfall, and low values of the lake and plantation indices. In contrast, class C includes the larger, flatter catchments with relatively low average annual rainfall and higher values of LAKE and PLTN. Class B has no particular distinguishing features, apart from being intermediate between classes A and C.

The next step in the evaluation of the benefits of clustering consisted of the development of separate MLPs to relate EVI parameters of the at-site flood frequency distributions to catchment characteristics for the drainage areas in each of the three clusters. In order to maintain comparability with the ANNs developed for different combinations of catchment characteristics on the combined date set, the 66 catchments that were selected for training and the 25 chosen for validation were employed for the same purposes for each of the three clusters. Hence, cluster A was trained on 20 sites and validated on 12, and cluster C was trained on 15 and validated on 4. The average RMSEs for the various clusters of catchments were found to be 119 m^3/s, 230 m3/s and 165 m^3/s for Clusters A, B and C respectively. The overall average performance of the individual cluster ANNs on the total set of 25 catchments was thus 180 m^3/s. This point is indicated on Figure 1 with an asterisk.

It is interesting to note how the performance of cluster B is inferior to clusters A and C for generalizing the validation sites. Cluster B is obviously less well differentiated than either cluster A or cluster C. The ´small´ and the ´large´ groups of catchments plainly have a clearer identity then the ´intermediate´ catchments. This latter cluster may perhaps be best described as a grouping of catchments that do not posses the well-defined properties of the other two clusters, and is worthy of more detailed investigation. However, the overall performance of the individual cluster ANNs shows a clear improvement over the ´unclustered´ ANNs. There is in fact an improvement of almost 31 % on the 6-variable ANN on the combined data set, and an event greater (44%) reduction on the RMSE for

The 4-variable regressin Equation (1). Although the problem of refining the intermediate clusterremains, there is a clear benefit to be obtained from clustering.

CONCLUSION

The annual maximum flood series compiled for the preparation of the FDMJS (1983) provide a typical example of the data sets available in developing countries for the purposes of regionalizing design flood estimates. The application of the widely-employed method of residuals to these data provided an equation for the index flood (MAF) whose ´worth´ in term of equivalent length of observations was less than 3 months. Moreover, mapping of the residuals between computed and observed MAFs did not reveal any consistent and widespread spatial patterns that would indicate the presence of sub-regional groupings. In contrast, use of an MLP-type of ANN with catchment characteristics as the input and flood frequency distribution parameters as the output was able to provide smaller RMSEs on both a verification data set and an independent set of sites for which only partial duration flood series were available.

When the characteristics of 92 catchments were subjected to a classification analysis using a Kohonen SOM, the existence of at least three distinct groupings of catchments was revealed. Two of the groupings, relating to small, steep catchments with more moderate AAR-values were hydrologically plausible, and consistent with the findings of a previous study on data (Hall and Minns, 1999). Since the Kohonen SOM may function as a sorting algorithm as well as a clustering routine according to the selection of parameters such as neighbourhood radius, the variation of catchments between classes is found to be monotonic. However, of the three classes of catchment identified, only the clusters for the small, steep areas and the large, flat drainage basins showed considerable improvement in their training and verification statistics. The intermediate group of catchments was only slightly better represented compared to the general ANN for all catchments. The composition of this intermediate cluster, especially the possibility of the presence of sub-clasters of hydrological relevance, must be the subject of an other investigation with this and other data sets.

REFERENCE

Benson, M. A., 1962. Factors influencing the occurrence of floods in a humid region of diverse terrain, *US Geol Survey, Watter-Supply Paper* 1580-B.

Burn, D. H., 1990. An appraisal of the “region of influence” approach to flood frequency analysis, *Hydrol. Sci. J.* 35: 149-165.

Dalrymple, T., 1960. Flood-frequency analysis, *US Geol Survey, Watter-Supply Paper* 1543-A.

FDMJS 1983, *Flood Design Manual for Java and Sumatra*, 2 Vols.

Gunter, B. N., 1974, The investigation of flood estimation procedures for Papua New Guinea, *Proc. Instn. Civ. Engrs. Part* 2 57: 635-650

Hall, M. J., Minns, A. W., 1998, Regional to flood frequency analysis using artificial neural networks, *Proc. Hydroinformatics ´98*, 3rd *Int. Conf. On Hydroinformatics*, Vol. 2, 759-763

Hall, M. J., Minns, A. W., 1999, The classification of hydrologically homogeneous regions, *Hydrol. Sci. J.* 44: 693-704.

Kohonen, T., 1995, *Self Organizing Maps*, Springer-Verlag, Berlin.

Linsley, R. K., 1967, The relation between rainfall and runoff, *J. Hydrol.* 5: 297-311.

Nathan, R. J., McMahon T.A., 1990, Identification of homogeneous regions for the purpose of regionalization, *J. Hydrol*. 121: 217-238.

Natural Environment Research Council, 1975, *Flood Studies Report*, 5 Vols.

Nouh, M. A., 1988, On the prediction of floo frequency in Saudi Arabia, *Proc. Instn. Civ. Ingrs. Part* 2 85: 121-144.

Wiltshire, S. E., 1985, Grouping basins for regionalflood frequence analysis, *Hydrol. Sci. J.* 30: 151-159.
World Bank,1993, *Water Resources Management. A World Bank Policy Paper.*