Journal of Geographic Information and Decision Analysis, vol. 2, no. 2, pp. 233 - 242, 1998

Rainfall Prediction Using Artificial Neural Networks

Sunyoung Lee
Department of Computer Science and Engineering, Pohang University of Science and Technology, San 31, Hyoja, Pohang, Kyungbuk 790-784, Korea
pulpiri@nova.postech.ac.kr

Sungzoon Cho
Department of Industrial Engineering, Seoul National University, San 56-1 Shinrimdong, Kwanakgu, Seoul 151-742, Korea
zoon@snu.ac.kr

Patrick M. Wong
Center for Petroleum Engineering, The University of New South Wales, Sydney NSW 2052, Australia
PM.Wong@unsw.edu.au

ABSTRACT **The spatial interpolation comparison 97 is concerned with predicting the daily rainfall at 367 locations based on the daily rainfall at nearby 100 locations in Switzerland. We propose a divide-and-conquer approach where the whole region is divided into four sub-areas and each is modeled with a different method. Predictions in two larger areas were made by RBF networks based on the locational information only. The two smaller areas were assumed to be implemented by the Orographic Effect which dictates that precipitation is proportional to elevation. Thus, predictions in these two areas were made using a simple linear regression model based on the elevation information only. Comparison with the observed data revealed that RBF networks produced good predictions while the linear models poor predictions. The relatively large prediction errors from the small areas seem to indicate that the Orographic Effect did not exist.**

KEYWORDS: **divide and conquer neural network, orographic effect, radial basis function network, universal function approximation.**

Contents

5. Conclusions
References

1. Neural networks

The brain consists of a large number of neurons, connected with each other by synapses. These networks of neurons are called neural networks, or natural neural networks. The artificial neural network (ANN) is a simplified mathematical model of a natural neural network. It is a directed graph where a vertex corresponds to a neuron and an edge to a synapse. Various ANN models have been proposed since its conception in the 1940's, but the multi-layer perceptron (MLP) and the radial basis function (RBF) network are the most widely used. These layered networks have been applied to various problems such as pattern recognition, prediction, and function approximation (Haykin 1994). Both models have been proven to be "universal function approximators" (Park and Sandberg 1991) which means that given enough data, the underlying function can be approximated with accuracy. In addition to these "existence" theorems, powerful and practical learning algorithms help find such networks. The most famous is backpropagation for MLP (Rumelhart *et al.* 1986).

RBF methods have their origins in techniques for performing the exact interpolation of a set of data points in a multi-dimensional space (Powell 1987). RBF mappings provide an interpolating function which passes exactly through every data point. If there is noise present on the data, the interpolating function which is smoother and which averages over the noise gives the best generalization. By introducing a number of modifications to the exact interpolation procedure we obtain the RBF neural network model (Broomhead and Lowe 1988, Moody and Darken 1989). Architecture of an RBF neural network is illustrated in Figure 1.

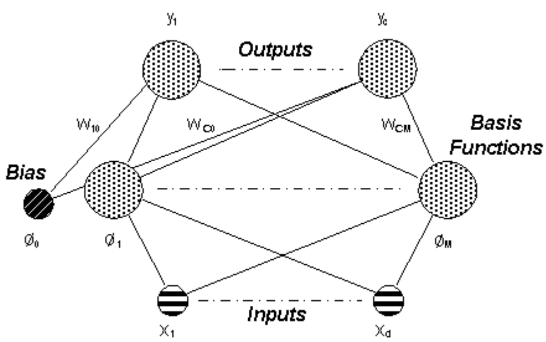

Figure 1 Architecture of a RBF neural network

ϕ_i is a basis function. Several forms of a basis function have been considered, the most common being the Gaussian,

$$\phi_i(x) = \exp\left(-\frac{(x-x_i)^2}{2\sigma^2}\right) \qquad (1)$$

where σ is a parameter whose value controls the smoothness properties of the interpolating function and x_i is the center of the basis function.
The outputs of the network are determined by linear combinations of these basis functions,

$$y_k(x) = \sum_{j=0}^{M} w_{kj}\phi_j(x) \tag{2}$$

where y_k denotes the k^{th} output value and w_{kj} represents the synaptic weight from the j^{th} basis function (or hidden node) to the k^{th} output.

RBF neural network training has the following two-stage procedure. In the first stage the input data set $\{x^n\}$ alone is used to determine the center locations of the basis functions. Using fixed basis functions, second-layer weights are found in the second phase of training.We can optimize the weights by minimizing the suitable error function. Generally one uses a sum-of-squares error function given by

$$E = \frac{1}{2}\sum_n \sum_k \{y_k(x^n) - t_k^n\}^2 \tag{3}$$

where t^n_k is the target value for output unit k when the network is presented with input vector x^n. Since the error function is a quadratic function of the weights, its minimum can be found in terms of the solution of the linear equations from the equations $y(x) = W\phi$ in matrix notation,

$$\Phi^T \Phi W^T = \Phi^T T \tag{4}$$

where $(W) = (w_{kj})$, $\phi = (\phi_j)$, $(T)_{nk} = t^n_k$, and $(\Phi)_{nj} = \phi_j(x^n)$.
The formal solution for the weights is then given by

$$W^T = \Phi^+ T \tag{5}$$

where Φ^+ denotes the pseudo-inverse of Φ . In the case of matrix instability, the equation can be solved by an iterative method.

2. Divide-and-conquer and orographic effect

Fitting all 100 training data to one single RBF network tends to result in poor interpolation performance. In many practical applications of a neural network, it has been shown that a "divide-and-conquer" approach results in improved performance (Cho *et al.* 1997). The classic computing procedure dictates that the given data set is divided into subsets and each subset is "conquered" or solved by a different network.

We made two assumptions regarding the rainfall prediction problem. First, general rainfall patterns can vary over different regions. Second, that these regions can be delineated by their elevation. Based on these assumptions, three different elevations were tried. At 500m, only two areas emerged in the resulting sectional view. Each area contained subareas of high altitude and low altitude. At 1000m, four areas were found. At 1500m, four areas were also found, but two of these areas had almost no observed data within them since they were very narrow. Thus, we decided to divide the whole region into four areas. The sectional view at 1000m is illustrated in Figure 3. Other intermediate values such as 750m and 1250m were not considered since it is difficult to predict which values will result in the best partition.

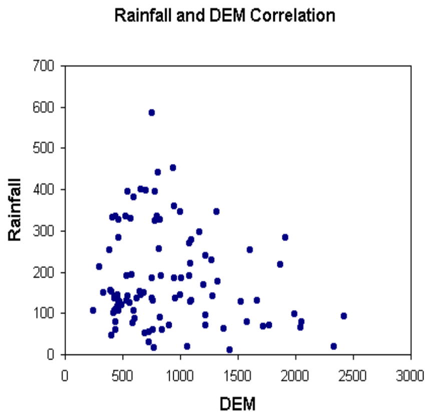

Figure 2 Correlation between rainfall and DEM

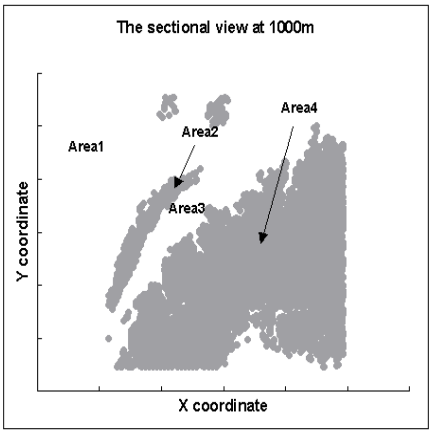

Figure 3 The sectional view at 1000m elevation

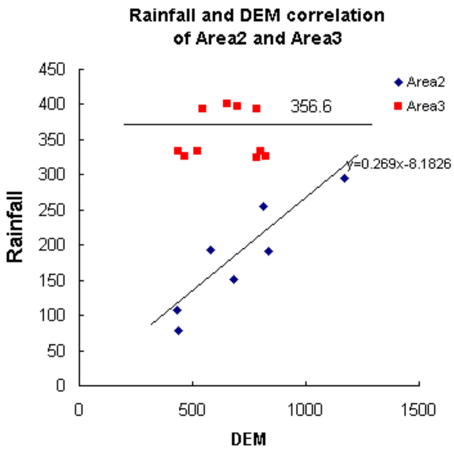

Figure 4 Correlation between rainfall and DEM

Table 1 Number of observed and estimated rainfall data points

	Number of observed rainfall data points	Number of estimated rainfall data point
Area 1	30	102
Area 2	7	32
Area 3	10	37
Area 4	53	196

Areas 1 and 4 are both large enough to contain sufficient observed/training data points. Thus, they were modeled by RBF networks. On the other hand, areas 2 and 3, are very small with insufficient observed data points. Figure 3 illustrates that area 2 is high and narrow ground, while area 3 is valley-like terrain. It has been known that on a mountain slope or a deep but narrow river valley, climatological precipitation typically increases with elevation (Alter 1919; Barrows 1933; Hibbert 1977; Schermerhorn 1967; Smith 1979; Spreen 1947). This phenomenon is commonly called the orographic effect.

Plotting the observed data in these regions against elevation (DEM) (Figure 4) seemed to confirm that the effect existed in areas 2 and 3. Therefore, we estimated the rainfall in areas 2 and 3 using a linear function of elevation.

3. Rainfall estimation method

Rainfall is estimated using the following methods in each area.

- Area 1 and Area 4: Two RBF neural networks are employed, one for each area. The inputs are x and y coordinates of the location. The output is the estimated rainfall. The Gaussian functions are used as basis functions of the hidden layer. Data points are not found to form clusters, thus a Gaussian hidden unit is placed at each data point. The smoothing parameter (σ value for Gaussian) is all set to 0.1. The number of basis functions is the same as the number of observed data in each area. The architecture of the RBF neural network is 2-30-1 for area 1 and 2-53-1 for area 4.
- Area 2 : Here we assume that the orographic effect exists. Thus, we estimate the rainfall based on the elevation, not on the location. The equation of the linear regression is $y = 0.269x - 8.1826$; where x is the elevation and y is the estimated rainfall.
- Area 3 : Here also the orographic effect is assumed to exist and we estimate the rainfall to be constant at 356.6.

The estimated rainfall values of areas 1 and 4 are found by minimizing the sum of square error between the observed and the estimated rainfall using leave-one-out cross validation. The leave-one-out cross validation mean square errors of areas 1 and 4 are 208.6 and 603.2, respectively. This leave-one-out cross validation mean square error value is used in the case of areas 1 and 4 as a measure of the accuracy of the predicted value.

4. Analysis of the estimation results

First, the trained RBF networks produced "negative" values for some cases (see Table 2). Postprocessing uniformly transformed all negative values to 10. The daily rainfall measurements are in $1/10^{th}$ of mm and values ranging from 1 to 10 are often not due to precipitation but to air condensation during the night. Choice of 10 over 1 or 0 was arbitrary and turned out to be a bad one.

Table 2 Estimated rainfall with negative value

ID	X	Y	Rainfall
115	-51974	-57968	-15.077194
121	-49688	-60211	-16.839237
139	-42749	-56929	-20.162487
143	-41935	-50262	-2.354461
176	-27345	-47010	-7.967682
356	63769	-109008	-22.173876
363	66081	-107871	-33.164387
366	67605	-105630	-45.731167
367	67556	-101183	-55.715054

The error values of the estimated rainfall are:

- Root Mean Square Errors (RMSE)
  Area 1 – 45.96
  Area 2 – 102.36
  Area 3 – 121.96
  Area 4 – 77.23
- RMSE of the whole dataset: 78.65
- Relative and Absolute Errors
- Relative Error – 0.46
- Absolute Error – 55.9

The prediction performance relating to areas 2 and 3 was not good. The orographic effect which we assumed to exist in these areas, apparently did not. If we had used the RBF networks for these areas instead, we would have obtained RMS errors of 86 and 72, for areas 2 and 3, respectively. The errors between observed and estimated rainfalls are illustrated in Figures 5 and 6. The correlation between observed rainfall and error is –0.34 and the number of over-estimated rainfall values is similar to that of under-estimated rainfall values.

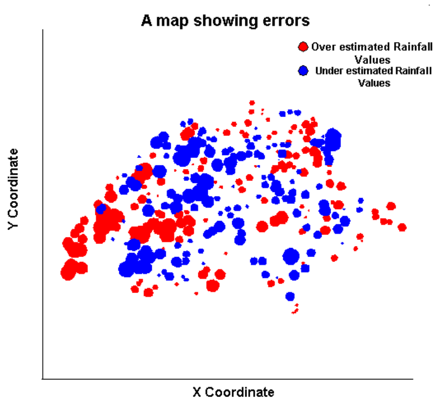

Figure 5 A map showing errors

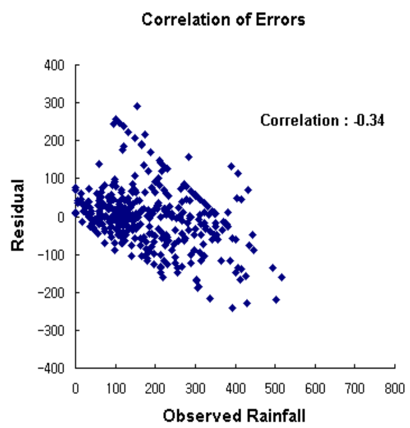

Figure 6 Correlation of errors

Table 3 Comparison between estimated and observed rainfall		
	Estimated Rainfall	**Observed Rainfall**
Min	3	0
Max	522	517
Mean	182	185
Median	148	162
Standard Deviation	112	113

The min, max, mean, median and standard deviation of estimated and observed rainfall are shown in Table 3. These statistics are more or less similar. The 10 lowest and highest values of estimated rainfall and observed rainfall are shown in Table 4. The 10 highest values of estimated rainfall are similar to those of the observed rainfall while the 10 lowest values of estimated rainfall are slightly different from those of observed rainfall. This was caused by the modification of the negative rainfall values into 10, instead of zero.

We plot the observed and estimated rainfall for each area, in Figures 7, 8, 9, and 10. Figure 11 illustrates a scatter diagram of estimated rainfall against observed rainfall for the whole region. The scatter diagrams, in Figures 7, 8, 9, and 10 confirm the prediction performance of RBF networks to be reasonable. However, in areas 2 and 3, the estimated rainfall values did not fit the observed rainfall values. We plot locations of the similar rainfall values with the same color, as shown in Figure 12.

Table 4 Ten highest values and ten lowest values

10 highest value		10 lowest value	
Estimated Rainfall	Observed Rainfall	Estimated Rainfall	Observed Rainfall
585	585	3	0
522	517	4	0
519	503	10	0
501	493	10	0
460	452	10	0
452	445	10	1
452	444	10	5
443	441	10	6
441	434	10	8
440	434	10	10

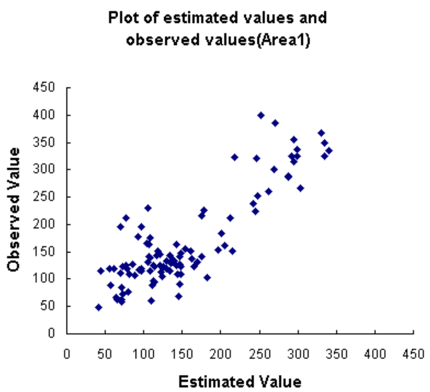

Figure 7 Plot of estimated values and observed values (Area 1)

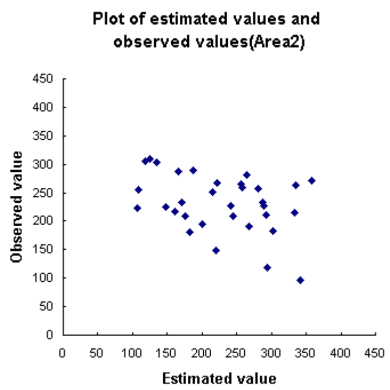

Figure 8 Plot of estimated values and observed value (Area 2)

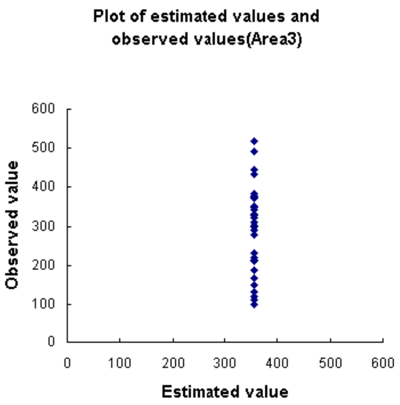

Figure 9 Plot of estimated values and observed value (Area 3)

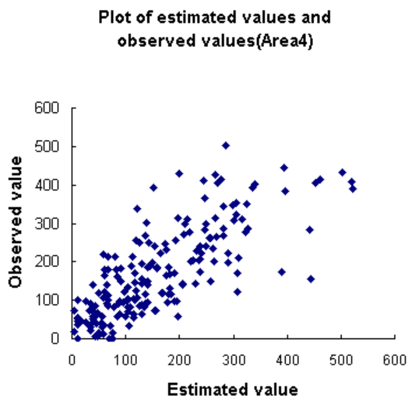

Figure 10 Plot of estimated values and observed value (Area 4)

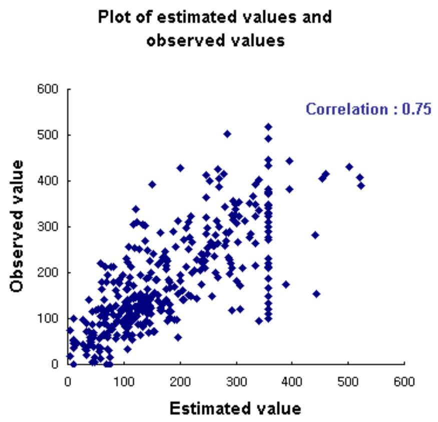

Figure 11 Plot of estimated values and observed values

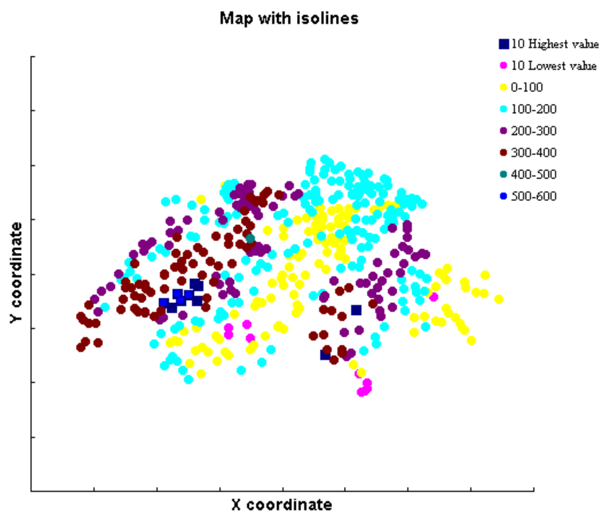

Figure 12 A map with isolines

5. Conclusions

For rainfall prediction, the "divide-and-conquer" approach was used to divide the whole region into four areas and the problem was solved separately based on the following assumptions: first, rainfall has a different pattern in different areas; second, rainfall pattern within smaller areas is continuous and smooth; third, orographic effect exists. Predictions of the two large areas using RBF networks were reasonably good. But, for the two small areas, predictions were not good since the orographic effect was not apparent. We could have obtained smaller errors of 86 and 72, rather than 102 and 121, respectively if we had employed RBF networks also for the small areas as well.

Nevertheless, the divide-and-conquer approach seemed the right choice. If we had fitted the whole region with a single RBF network, we would have obtained a larger RMS error of 87. If we had also used the elevation variable for the prediction, the overall RMS error would have decreased to 67. The MLP model, another universal approximator, did not prove to be as effective as RBF for this problem.

The proposed method does not require pre-modeling or any other tools. Since we used the same number of basis functions as observed data points, data clustering was not necessary. The weight values from the hidden layer were easily found using the linear matrix inversion technique. Our proposed method is suitable for emergency conditions, as well as long term management of contaminated regions.

References

Alter, J. C. (1919) Normal precipitation in Utah. *Monthly Weather Review* 47; 633-636.

Barrows, H. K. (1933) Precipitation and runoff and altitude relations for the Connecticut River. *Trans. Am. Geophys. Union*, 14, pp. 396-406.

Broomhead, D. S. and Lowe, D. (1988) Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2; 321-335.

Cho, S. , Cho, Y. and Yoon, S. (1997) Reliable roll force prediction in cold mill using multiple neural networks. *IEEE Transactions on Neural Networks*, 8(4); 874-882.

Haykin. (1994) *Neural Networks*, Englewood, N.J.: Prentice-Hall.

Hibbert, A. R. (1977) Distribution of precipitation on rugged terrain in central Arizona. *Hydrol. Water Resour. Ariz. Southwest*, 7, pp. 163-173.

Moody, J. and Darken, C. J.(1989) Fast learning in networks of locally-tuned processing units. *Neural Computations* 1 (2), pp. 281-294.

Park, J. and Sandberg, I (1991) Universal approximation using radial basis function networks. *Neural Computation*, 3(2), 246-257.

Powell, M. J. D. (1987) Radial basis functions for multivariable interpolation: a review. In J. C. Mason, and M. G. Cox (Eds.), *Algorithms for Approximation*, pp. 143- 167. Oxford: Clarendon Press.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) Learning internal representations by error propagation. In D.E. Rumelhart, J. L. McClelland, and the PDP Research Group (Eds.), *Parallel Distributed Processing : Explorations in the Microstructure of Cognition,* Vol. 1: Foundations, pp. 318- 362.

Schermerhorn, V. P. (1967) Relations between topography and annual precipitation in western Oregon and Washington. *Water Resources Research* 3; 707-711.

Smith, R. B. (1979) The influence of mountains on the atmosphere, *Advances in Geophysics*, Vol. 21, New York: Academic Press, pp. 87-230.
Spreen, W. C. (1947) A determination of the effect of topography upon precipitation. *Trans. Am. Geophys. Union*, 28; 285-290.

JGIDA vol. 2, no. 2
JGIDA Home