Data Distributions on Performance of Neural Networks for Two Year Peak Stream Discharges

Ranjan S. Muttiah, Assistant Professor
Texas Agricultural Experiment Station
808 E. Blackland Road
Temple, Texas 76502
email: muttiah@brcsun0.tamu.edu

Abstract

The impact of the input and output probability distributions on the performance of neural networks to forecast two year peak stream flow (cubic meters per second) is examined for two major river basins of the US. The neural network input consisted of drainage area (square kilometers) and elevation (meters). When data are normally distributed, the neural networks predict much better than when the data are non-normal and have larger tails in their distributions.

Keywords: Neural networks, data distributions, performance

Introduction

Cascade Correlation Neural networks were found to be viable predictors of the two year peak discharges for many parts of the US (Muttiah et al., 1996). The cascade correlation network dynamically adds hidden neurons as training progresses, and each neuron uses sigmoid transfer functions (Fahlman and Lebiere, 1990). The neural networks were trained on data that had been grouped by the major two digit river basins of the US (Seaber et al., 1994). The input data to neural networks consisted of drainage area (da), and elevation (elev), and the output data consisted of the observed two year peak stream flow discharges (q2). The highest correlations of 0.95 with measured two year peak discharges were obtained for the river basins in California (CA), and the lowest correlations (0.73) with measured data were obtained for the Souris-Red Rainy (SRR) river basin located in North Dakota and Minnesota. The reasons for the variations in correlations were thought to be the variability of the input data (elevation and drainage area) in the river basins, and the number of training samples used for the neural networks. We explore whether the probability data distributions of the input and output data was an additional factor in neural network performance. The results presented here could be used by

modelers to determine how well test data will be predicted by neural networks.

Previous theoretical work on the neural network performance has suggested that the number of training samples has to be just adequate for the neural network interpolation and extrapolation so that the "noise" in the data isn't learned, and that the input data distributions have to be similar to those of the squashing transfer functions used by the individual neurons in the network (Baum and Haussler, 1989; White 1989). For example, if the input data were normally distributed then Gaussian squashing functions (as well as sigmoid squashing functions since two sigmoids add to one Gaussian squashing function) would adequately forecast test data.

Objective

Determine whether the input and output data distributions had an impact on neural network performance for the two year peak stream discharge predictions.

Method

The input data from each of the river basins (CA, and SRR) were segmented into vectors and read into the S-PLUS statistical package using the scan() routine (S-PLUS User's Manual, 1995). Then the input vectors and output vectors were displayed and plotted out using the qqnorm(), qqline(), and qqplot() routines. The qqnorm() routine plots the quantile probability plot of the data, and qqline() plots a straight line showing where the data should be were it to be normally distributed, and the qqplot() routine plots the quantiles of two data vectors against each other. If the qqplot were a straight line, then that would signify that the data were distributed similarly. The Wilk's ratio that shows linearly independence and lack of cross-correlation of the data were also generated (1.0 for complete independence).

Some data on the two river basins were as follows:

CA:

Total number of data:	741
Wilk's ratio:	0.99
Training data (train n):	600
Test data (test n):	141

SRR:

Total number of data:	254
Wilk's ratio:	0.43
Train n :	150
Test n:	104

The number of data for the other two digit river basins for the US ranged from a low of 122 for the lower Mississippi river basin to a high of 1786 for the Missouri river basin. The Wilk's ratios ranged from a low of 0.27 for Arkansas to a high of 0.99 for California and South Atlantic (the southern states of US). The correlations against measured values for each of these river basins were as follows:

	corr.	Train n	Test n
Lower Mississippi:	0.88	75	47
Missouri:	0.79	1500	417
Arkansas:	0.85	350	104
South Atlantic:	0.94	1000	437

The lower correlation with measured values for Missouri may be due to the wide variation in topography, and climatic factors for that large a region (nearly one fifth the size of US involving six states).

Results

Figure 1 shows the quantile plot of drainage area for the Souris-Red Rainy. The plots shows the data to be quite non-normal with large tails. Similarly for the elevation input distribution. The quantile plot of two year peak flow is more normally distributed than the inputs but still has large tails in the distribution. When drainage area is plotted against discharge (Figure 2), the line is linear for low flows and drainage areas, but becomes nonlinear for high flows and drainage areas ($R^2 = 0.613$). When elevation is ploted against two year peak flow, non-linearity occurs at about 760 meters ($R^2 = 0.01346$).
Figure 3 shows the quantile plot of drainage area for California. The majority of the data are normal, and there is an excessive tail at very high areas (i.e., it is lognormally distributed). The quantile plot for elevation is more normally distributed than that for SRR. When the quantile-quantile plot is examined for drainage and peak flow, an R^2 of 0.177 is obtained (Figure 4). The q-q plot for elevation versus two year peak flows show a non-linear relationship developing at above 1,800 meters ($R^2 = 0.01733$).

Conclusions

When the elevation and drainage areas (input variables) are each individually closer to a normal Gaussian distribution the better the forecasts are for two year peak discharge (output variable) using neural networks with sigmoid transfer functions. The quantile relationship between input and output variables does not suggest future performance of neural networks.

References

Fahlman, S.E., and C. Lebiere. 1990. The Cascade-Correlation Architecture. CMU-CS-90-100, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvannia.

Baum, E.B., and D. Haussler. 1989. What Size Net Gives Valid Generalization ? NIPS I. Ed. D.S. Touretzky. Morgan Publishers, 2929 Campus Drive, San Mateo, CA 94403. pp. 81-90.

Muttiah, R.S., R. Srinivasan, and P.M. Allen. 1996. Prediction of Two Year Peak Stream Discharge using Neural Networks. Water Resources Bulletin (in review). American Water Resources Association, 5410 Grosvenor Lane, Suite 220, Bethesda, MD 20814-2192.

Seaber P.R., F.P. Kapinos, and G.L. Knapp. 1994. Hydrologic Unit Maps. U.S. Geological Survey Water-Supply Paper 2294. USGS, Books and Open-File Reports, Federal Center, Box 25425, Denver, CO 80225, USA.

S-PLUS Users' Manual. 1993. Version 3.2, MathSoft, Inc. 1700 Westlake Ave. N, Suite 500, Seattle, Washington 98109, USA.

White, H. 1989. Learning in Artificial Neural Networks: A Statistical Perspective. Neural Computation. B1, 425-464. MIT Press, Cambridge, MA.

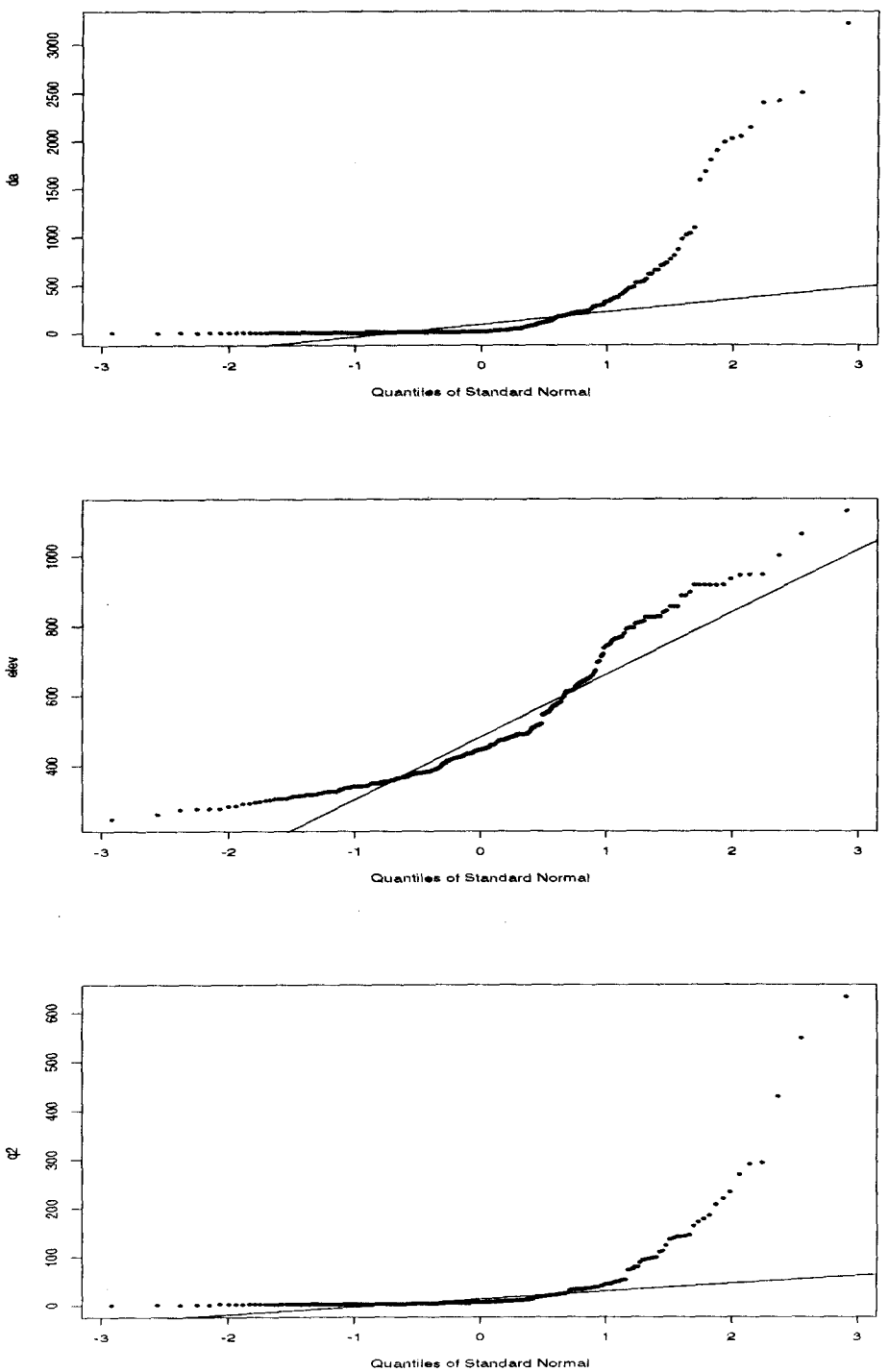
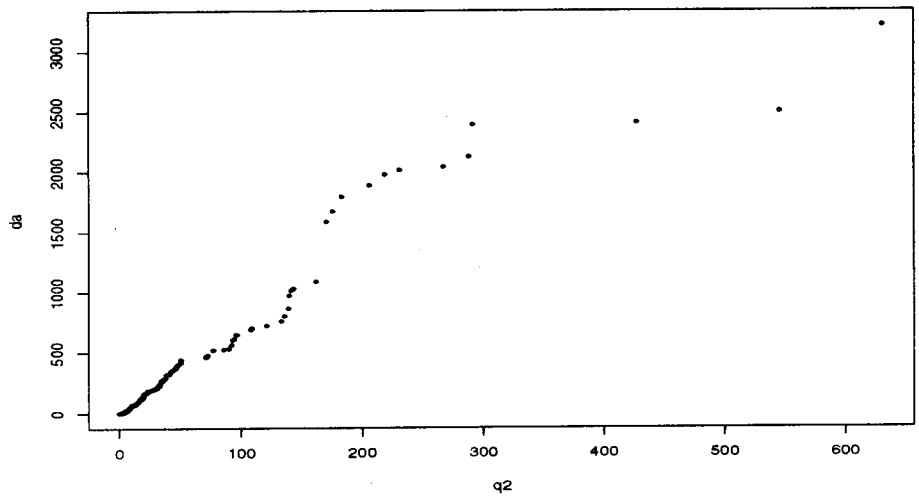


Figure 1. Probability distributions of the input (da, elev) and output (q2) variables for Souris-Red Rainy.



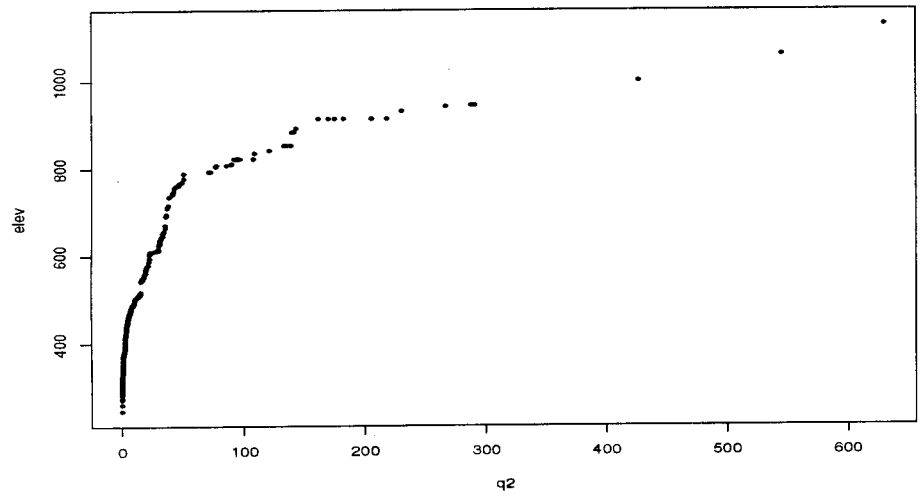


Figure 2. Quantile-Quantile plots for drainage area and elevation versus two year flow for SRR.

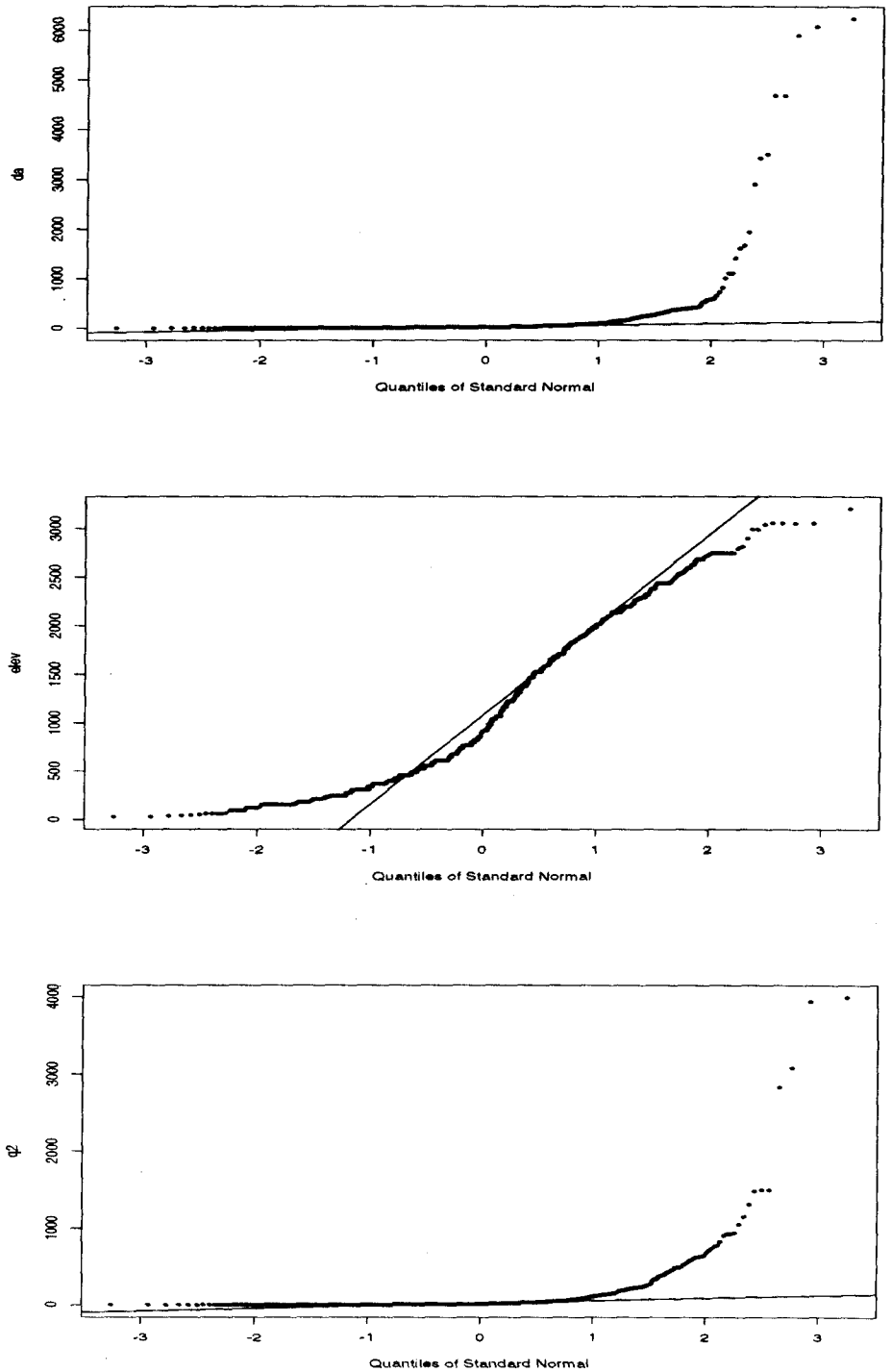
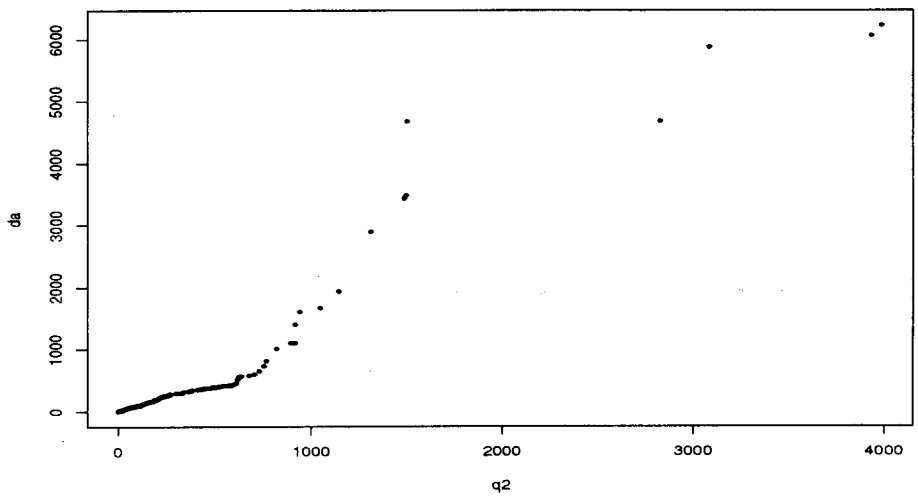


Figure 3. Probability distributions of the input (da, elev) and output (q2) variables for California.



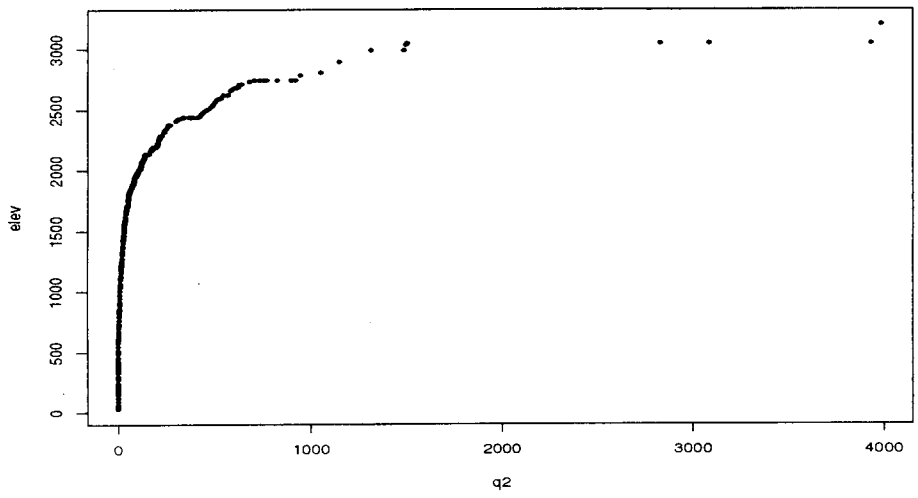


Figure 4. Quantile-Quantile plots for drainage area and elevation versus two year flow for California.