USING THE RADIAL BASIS FUNCTION TO PREDICT THE QUALITY OF MISSISSIPPI GULF COAST OYSTER REEFS

**Roger King, *William C. Powell, and *Joseph Sherrard

*Department of Civil Engineering
**Department of Electrical Engineering
Mississippi State University

INTRODUCTION

The Mississippi Gulf Coast oyster reefs are divided into eight areas. Each area is classified as approved, conditionally approved, restricted, or conditionally restricted based on a sanitary survey as designated by the National Shellfish Sanitation Program (NSSP) Manual of Operations developed by the U.S. Food and Drug Administration (FDA). Areas are monitored at their respective sampling stations for fecal contamination on a continual basis. These stations are used to classify zones within each area. A zone is closed down when its respective sampling station indicates through lab testing that fecal material, pathogenic microorganisms, and/or poisonous and deleterious substances are present within the zone, and the zone cannot be classified as approved. The determination that the approved classification standards are met is based upon a minimum of 15 samples collected from the zone's respective station. The zone is classified as approved when the fecal coliform median or geometric mean Most Probable Number (MPN) of the water does not exceed 14 per 100 ml and not more than 10 percent of the samples exceed an MPN of 43 per 100 ml (Food and Drug Administration 1992).

A variety of meteorological, hydrographic, and geographic factors affect the distribution of pollutants over each station. Factors include rainfall, river stage, change in river stage, salinity, water temperature, wind speed and direction, and tidal stage. Since 1989, sampling station bacterial results have been documented along with the existing factors present at the time of sampling (Ladner 1994). This existing data can now be used to predict the MPN of the water around each sampling station using the Radial Basis Function.

For purposes of this paper, only one of the eight classified areas is studied. This area, AREA 1, has a total of 13 stations and 479 valid input data patterns.

REASONING THE USE OF THE RADIAL BASIS FUNCTION

Five primary factors will be examined for the prediction of the oyster bed water quality: rainfall, river stage, rise in river stage, sampling station, and the month in which the sample was taken. Only these factors are used because their data are easily and readily obtainable when they are needed. Also, these factors have the biggest impact on the MPN. The other factors are not included because they either are determined by sampling or are not readily obtainable when they are needed, and their impacts upon the MPN are secondary at best.

The Radial Basis Function (RBF) is applicable because an exact function is not possible without incorporating the other secondary factors into the data set. Also, there may be other factors which may not yet be known. The RBF's ability to generalize is needed to take into account the absence of those factors and the "fuzziness" of the accuracy of the data.

INPUT PARAMETERS

Non-point source pollution is a major contributor of coliform organisms found in coastal waters. During heavy rainfall, the coliforms on the ground are flushed into rivers which empty into the Gulf of Mexico. Storm water also empties as runoff into the Gulf. During and after these events, heavy concentrations of fecal coliform enter the Gulf Coast waters, polluting the oyster beds and causing a rise in the MPN. These flushes of fecal coliform are characterized by a rise in river stage, a high river stage, and heavy rainfall. This provides the basis for which the input parameters were created. To be specific, the input parameters used and their reasoning are as follows:

River stage. Gives an indication of how much rainfall from the river watershed area has emptied into the Pearl River. Therefore, a high river stage indicates high amounts of rain in the watershed area, and a high MPN may be

expected, though such is not always the case. A rise in river stage seems to be a more critical indicator.

Rise in river stage. Shows the extent of the initial flush of rainfall into the river. This initial flush washes the majority of coliforms into the river. Therefore, a high rise in river stage indicates a high flush of coliforms into the river, which in turn empties into the Gulf, and a high MPN may be expected.

Days of sustained or declining river stage. Indicates whether or not any minor rises in river stage have occurred since the last significant rise in river stage. This will give an indication of any smaller amount of coliforms that may have been emptied into the Gulf in addition to the last major flush.

Station number. Shows the susceptibility of a specific station to being contaminated by fecal coliforms. This is also used as an indicator of the extent of contamination that can be expected at nearby stations.

Month. Documents the month in which a sample was taken. The reasoning is that colder months are more susceptible to contamination than summer months. In the summer months, a greater percentage of rainfall is either absorbed by the dry soil or evaporated by the sun. Increased plant photosynthesis also takes up much water. Therefore, less rain is able to flush into the river, and the river stage generally stays low as a result. In the colder months, however, plant life is generally inactive and little evaporation occurs. Thus, the soil becomes susceptible to saturation. With the soil near saturation point, and with little evaporation taking place, most of the rainfall is converted to runoff. Thus, rainfall data are used for up to five days prior to sampling. This gives an indication of not only how much rainfall has occurred over the past five days, but also the degree of saturation of the soil. Four rainfall data collection points are averaged to help filter out inputs from spot showers and give a better indication of how much rainfall has actually occurred.

HOW THE RBF WORKS

Radial basis function networks were selected for this research because of their ability to rapidly train and because they do not exhibit some of the local minima problems associated with back propagation networks. They also can serve as universal approximators, given that the network has enough nodes in the hidden layer.

Although the RBF architecture looks similar to the back propagation network architecture (i.e., input layer, hidden layer, output layer), it is fundamentally different in its training paradigm. Rather than starting with random values on all of the interconnections between the layers, the weights for the interconnections between the input layer and the hidden layer are set to values that give a particular desired response. Each node in the hidden layer is designed so that its response covers a particular portion of the input space, usually with a Gaussian exponential activation function. This can be visualized as a bump in the n-dimensional input space. The "centers" of these bumps represent exemplars of the data. These exemplars are found via the use of a clustering algorithm. The coordinates of the exemplar in the n-dimensional input space are then set as the weights for the input to hidden layer interconnections. The clustering of the data for the hidden layer simplifies the training process and results in the rapid training achievable with RBF.

The Radial Basis Function is easily represented by a three layer feedforward neural network which utilizes the clustering of data. The input layer consists of n units (in this case n = 11) which represent the elements of a vector x. The hidden layer consists of K units, which is selected as the number of clusters desired by the user.

The Radial Basis Function is derived from the theory of approximation. Given N pairs (x_i, y_i), a function f is searched for in the form (Zell et al 1994):

$$f(x) = \sum_{1}^{K} C_i h(|x - t_i|)$$

where h is the radial basis function and t_i are the K centers which are selected. The coefficients c_i are unknown and have to be computed while x_i and t_i are the elements of an n-dimensional vector space. h is applied to the Euclidian distance between each center t_i and the given argument x. In this research, the Gaussian function is used. The values of x which are equal to a center t yield an output value of 1.0 for the function h, while the output becomes almost zero for larger distances.

The hidden units compute the Euclidian distance between the input pattern and the vector which is represented by the links leading to this unit. By applying the euclidian distance to the function h, the activation of the hidden units is computed.

The output nodes typically perform a summation of the hidden layer node outputs to achieve the desired output. Therefore, the leading to this unit. By applying the Euclidian distance to the function h, the activation of the hidden units is computed.

The output nodes typically perform a summation of the hidden layer node outputs to achieve the desired output. Therefore, the training of the weights between the hidden layer and the output layer results in the training of a single layer linear network. Again, this improves the training time. A supervised learning paradigm is used to output layer weights based upon the desired input/output response.

PROCEDURE

In the first step, all input parameters are first normalized from zero to one as required by the SNNS radial basis function system. After a fully connected three-layer feedforward network is constructed, the link weights between the input and hidden layer are then set using the procedure RBF_Weights_Kohonen in the SNNS system so that the center vectors (represented by the link weights) form a subset of the teaching pattern. The bias is then checked and set.

In the next step, the initialization procedure is begun using the function RBF_Weights. After initialization, the training of the network begins. The three learning rates (centers, bias, and weights) are examined for sensitivity and set to the appropriate values. The learning function is run through cycles until the error has decreased to an acceptable level. The results can now be checked.

RESULTS

Once operation of the SNNS system has been understood and the network is initialized, learning time takes about 6 hours provided no mistake has been made in setting any of the three learning rates. With the input of a momentum term and a maximum error term, the learning time can be cut down to about 2 hours, but the learning rates become more sensitive and their selection must be careful.

The error was calculated as the sum of squared errors of all the output units. While the error decreased rapidly at first, it leveled out before reaching an acceptable level, showing problems with the data input which will be discussed later in this section.

The MPN values of the input ranged from 1 to 1600, showing a wide variation of outputs that may be encountered. After these values were normalized, the critical point at which a station (or zone) can no longer be classified as approved occurred at an output less than 0.009. Thus, an output less than 0.009 predicts that the station should be classified as approved while an output greater than 0.009 predicts that the station should not be approved and no oystering should be allowed.

Based on this all or nothing condition, the network was shown to be only 78% effective in predicting the true classification as compared with the actual output data. Therefore, a deviation parameter was developed to encompass those network output values which were more susceptible to being incorrect. This deviation value ranged from 0.006 to 0.020. The outputs that fell within this range were then classified as questionable and not considered in calculating the effectiveness. With these suspect outputs removed, the effectiveness of the prediction increased to 87%. While this value is an improvement, it can be further refined. Therefore, another attempt to increase the effectiveness was made. Since 5 of the 13 stations were no longer being used and no data were taken from them in over three years, they were discounted. A total of 132 out 479 network outputs were removed from consideration. This increased the overall effectiveness of the network from 78% to 84%. With the deviation parameter included, the effectiveness increased to 93%. This value is a further improvement, but it is still below the effectiveness expected.

With the inclusion of a deviation parameter, the network outputs can now be grouped. Outputs that range below 0.006 can be classified as good water quality. Outputs that range above 0.020 can be classified as poor water quality; and outputs that range between 0.006 and 0.020 can be classified as questionable water quality.

DISCUSSION

Even though the effectiveness of the network never lived up to expectations, these are just preliminary results. The error of the network suggests that other factors that are not readily seen are affecting the MPN of the output data set and must be considered in some way.

One indication of problems found in the data set is the presence of a high MPN (as high as 540 compared with the approved value of 14) with no adverse conditions occurring, suggesting that some other source other than the river or rainfall runoff was responsible for the contamination. This output obviously affects the network's learning ability. There are many possible sources of this contamination which are hard to predict. Construction on the shoreline may be responsible, especially with gambling becoming so popular along the coast. Boats may be another possible source. Many boats are equipped with on-board toilets which are flushed directly into the Gulf, spreading contamination. A heavy contamination may occur in the presence of many such boats, such as in a boat race or boat party. Summer activity can also affect those stations located close to shore.

Another indication of problems is the presence of a low MPN in the presence of an adverse condition, but this occurrence is fairly rare. The most probable explanation would be to attribute it to human error in recording the data. This can also help account for the presence of high MPN with no adverse conditions occurring.

CONCLUSIONS

The radial basis function still has merit even though the network never reached the effectiveness hoped for. A variety of unpredictable sources of contamination may well be responsible for the network's error, and further research will continue in the coming months to filter out these sources of error to provide a more accurate reflection on the actual results the radial basis function is capable of performing. With the effectiveness already reaching 93% without considering these factors, future research will undoubtedly produce the desired results.

REFERENCES

Food and Drug Administration. 1992. National shellfish sanitation program manual of operations. (revision).

Ladner, Cornell (Dr.). 1994. Engineer at the Mississippi Bureau of Marine Resources. Personal communication.

Zell, Andreas, Niels Mache, Ralf Hubner, Gunter Mamier, Michael Vogt, Kai-Uwe Herrmann, Michael Schmalzl, Tilman Sommer, Artemis Hatzigeorgiou, Sven Doring, and Dietmar Posselt. 1994. Stuttgart neural network simulator user manual, Version 3.1.