PREDICTION OF SALINITY IN SAN FRANCISCO BAY DELTA USING NEURAL NETWORK

T. RAJKUMAR[1] and MICHAEL L. JOHNSON

Metabyte Networks, Fremont, California 94538 USA
John Muir Institute of the Environment,
University of California, Davis, California 95616 USA

Abstract

We used an artificial neural network approach to predict salinity intrusion, a complex problem in the Sacramento-San Joaquin River delta in California. The inputs are comprised of the flows from the Sacramento and San Joaquin rivers, and stage data from different locations around the bay and delta region. Different neural network architectures and training algorithms were applied to this problem to find the optimal architecture to satisfy all possible scenarios. Out of all training algorithms tested, the back propagation method using the Levenberg-Marquardt algorithm was the best predictor of salinity intrusion.

The neural network was composed of three layers with the hidden layer of neurons consisting of three times the number of input neurons. Predicted salinities were within ten percent of the actual salinity at Carquinez strait (RSAC054) measured for two periods of time, April 1997 and August 1998. Two selected management scenarios consisting of increased pumping at the federal and state water projects were evaluated to determine the resulting change in salinity at Carquinez strait. Increased pumping by fifty percent resulted in an increase in salinity of twenty percent. The speed with which these predictions can be evaluated indicate that a neural network approach could be used to evaluate a large number of potential management scenarios to determine their general effects on salinity intrusion into the delta.

Keywords

Neural network, Non-linear modeling, Salinity prediction

1 Introduction

The Sacramento-San Joaquin river delta is an integral part of California's water supply network (Figure 1). Water projects divert water from the delta to meet the needs of approximately two-thirds of the state's population. In addition to these exports, there are about 2000 local irrigation diversions located within the delta. Most of the remaining water constitutes the final delta outflow into San Francisco Bay. Saline water from the bay is normally kept from intruding into the delta by these freshwater outflows. Saltwater intrusion into the delta is a major water quality problem that affects both wildlife and humans.

The initial problem in managing salt-water intrusion is to evaluate the salinity at a given location in the delta at any point of time. Traditionally salinity is measured using monitoring instruments installed at various locations within the bay-delta system. These data can be used to predict future conditions and evaluate the management alternatives. The California Department of Water Resources Delta Simulation Model (DWRDSM) [1] is used for this type of analysis. DWRDSM [2,12] is an unsteady one-dimensional hydrodynamic and salt transport model. It can provide estimates of salinity at almost any location within the bay-delta system. However, like many numerical models, when simulating the dynamics a large complex system, the processing time can be extremely long which makes it difficult to evaluate a large number of management scenarios in a short time. Statistically based models of the flow and salinity have been tried and found lacking [15]; classical time series analysis [14] is often linear or requires transformation to a stationary time series which

0-7803-7087-2/01/$10.00 © 2001 IEEE

renders the resulting model unsuitable, since the salinity intrusion problem is a complex non-linear problem. Consequently, there remains a need for a reliable and fast method to predict salinity in the Sacramento-San Joaquin river delta system. Our goal is to develop a tool for reliability and rapidly predicting salinity in the Sacramento-San Joaquin river delta system. Our criteria for reliability include prediction of salinity to an accuracy of ±10% at hourly intervals for any time period in the year.

Our specific objective is to predict the salinity and stage at Carquinez strait for hourly intervals. We selected the Carquinez strait because it lies at the western edge of Suisun bay and is generally near the location of biologically productive salt-fresh water mixing zone. Also, monitoring data are available for Carquinez strait, which allows model development and testing.

2 ANN development

The success of the neural network [6,7,8,11,13] depends greatly on defining the influencing parameters for the problem. Figure 1 represents the outline of the bay-delta region and details of stations are provided in Appendix I. Hourly data for several locations were obtained from the Interagency Ecological Program. (Interagency ecological website http://www.iep.ca.gov)[5]. The data set consisted of 20 input variables and one output variable and the data represents hourly collected data for all variables around 20 different locations in bay-delta. The size of the data matrix can be represented as 365x24 rows and 21 columns. Data used are from the years 1996-1998. 1996 and 1997 are considered to be the training data set and 1998 constitutes the prediction data set. Flow values are given in cubic feet per second (cfs), Salinity is measured as electrical conductivity in micro mhos per centimeter (mmhos/cm), and stage in feet (ft). The meteorological variables like wind speed are expressed in m/sec and barometric pressure as millibars. Any missing value for any variable in the data resulted in that hour's data being omitted from the data set. The input and output values are scaled to range [0.1 0.9], to insure that the output will lie in the output region of the nonlinear sigmoid transfer function. Presentable variable values lie in between 0.1 and 0.9 (0.1 and 0.9 inclusive).

3 Algorithm selection

There are many kinds of neural networks are available for use. Nobody knows exactly how many. New ones (or at least variations of old ones) are invented every week. The critical issue in developing a neural network is generalization: How well will the network make predictions for cases that are not in the domain of the training set? Neural networks, like other flexible nonlinear estimation methods such as kernel regression and smoothing splines, can suffer from either underfitting or overfitting. A network that is not sufficiently complex can fail to detect fully the signal in a complicated data set, leading to underfitting. A network that is too complex may fit the noise, not just the signal, leading to overfitting. Overfitting is especially dangerous because it can easily lead to predictions that are far beyond the range of the training data with many of the common types of neural networks. Overfitting can also produce wild predictions in multilayer perceptrons even with noise-free data. The best way to avoid overfitting is to use *lots* of training data. If we have at least 30 times as many training cases as there are weights in the network, we are unlikely to suffer from much overfitting, although we may get some slight overfitting no matter how large the training set is. For noise-free data, 5 times as many training cases as weights may be sufficient. But we can't arbitrarily reduce the number of weights for fear of underfitting. To overcome all the potential problems with overfitting and underfitting, we conducted seven different sets of analyses to establish the appropriate training algorithm, the best architecture, and select the best variables to include in the artificial neural network.

The first analysis was used to determine the most efficient training algorithm for this type of problem. In the first analysis, we considered only the locations on the upstream side of the Carquinez strait (Figure 1) to evaluate the basic training algorithm required for the neural network. Fourteen different training algorithms were tested (Bayesian regulation, BFCGS Quasi-Newton, Conjugate gradient back propagation with Powell beale restarts, conjugate gradient, conjugate gradient with polka ribiere updates, gradient descent, Gradient descent with adaptive learning, Gradient descent with momentum back propagation, gradient descent with momentum and learning, Levenberg-marquardt back propagation, One step secant back propagation,

RPROP back propagation, scaled conjugate gradient, By weight and bias 1-vector at a time training function, and an elman recurrent neural network). Of the fourteen algorithms, the Levenberg-Marquardt back propagation [10] and Bayesian regulation back propagation [9] were the most efficient algorithms and resulted in the least amount of overfitting and underfitting. Some data points were under or over fitted in this experiment. To resolve the under or over fitting we further experimented using different data sets and architecture.

In the initial analysis we considered the upstream data points for carquinez strait. Upstream data alone were not sufficient to calibrate efficiently the salinity intrusion analysis, which is caused partly by the tidal flux from the Pacific Ocean. We further enhanced the model with downstream inputs as well moon phase (lunar phase the causes tidal effect). The question here is to analyze which are all the key parameters, which contribute significant influence on salinity intrusion.

4 Architecture and Variable selections

The next steps were used to analyze the input variable selection and the network architecture based on the Levenberg-Marquardt and Bayesian regulation learning techniques. In the second analysis, tidal effect was introduced by adding the downstream locations. The major input of salinity is from ocean water transported into the delta by the tides. Minor amounts of salt are added and removed by agricultural activity within the delta. For example an average tidal flow through the Golden Gate is 2,300,00 cfs and the average delta outflow ranges from 600-32000 cfs. In the second analysis, the Bayesian regulation back propagation technique did not reach the error of tolerance and it was eliminated in further analysis.

In the third analysis some farther locations were eliminated and locations closer to the Carquinez strait were added. The Levenberg-Marquardt algorithm produced reasonably good results. Still some data points were under/over fitted. In the fourth analysis, we included meteorological variables like wind speed and barometric pressure from two location in the bay-delta region. Including meteorological data did not result in any reasonable improvements over the earlier analysis. Up to the fourth analysis, the neural network architecture considered was a three-layer neural network with equal number of input and hidden layer neurons.

In the fifth analysis, we focused on the tidal flux towards the delta by incorporating the phase of the moon. Every day two high tides and two low tides occur in the San Francisco bay area. In this analysis the meteorological variables were omitted and instead the moon phase illumination is incorporated. The percent of the Moon's surface illuminated is a more refined quantitative description of the Moon's appearance than is the phase. This variable is used similarly to a bias variable in the neural network. The moon's surface illumination varies from full moon to new moon in a continuous manner throughout the year. This pattern is altered slightly once in 19 years. The change in the moon's illumination in 19 years is very small compared to the prediction of salinity and it did not have an impact on the prediction of salinity. In the fifth analysis various neural network architectures were considered. We found that a three layer neural network with three hidden neurons for each input neuron produced excellent results. Flows and salinity were simulated well with very little overfitting or underfitting.

The sixth analysis was a combination of the third and fifth analyses. In the sixth analysis, we further extended the parameters with various neural network architectures, which resulted in a reasonably good fit. The Levenberg algorithm did not increase the performance noticeably once the number of neurons in the network exceeded more than 100. The gradient descent algorithm [3,4] was generally very slow, because it requires small learning rates for stable learning. Networks are also sensitive to the number of hidden neurons in the hidden layer. Too few neurons can lead to under fitting. Too many neurons can contribute to over fitting, in which all training points are well fit, but the fitting curve takes wild oscillations between these points. In the final analysis, we did not increase the number of neurons beyond 100.

In the final analysis, we considered pumping from the three major pumping stations located in the delta. This analysis focused on the impact of pumping and how the salinity intrudes into the delta region. These pumps remove water from the delta and send it south to the Central Valley of California and Los Angeles. A large volume of water is drawn out of the delta, which allows

the saltwater intrusion. When we included pumping volume as a variable, it produced an excellent result in predicting the salinity. Pumping has a high impact on the delta outflow, which pushes the saltwater further to the east in the delta. There are three major pumping stations located in the San Joaquin River in the delta, which draw a considerable amount of water (i.e. 11328 cfs max. each pumping station). The input variables considered were the combination of variables used in the fifth analysis and the draw down from the three major pumping stations. The total number of input variables considered for this experiment was 15 with one output variable (salinity). From previous analyses we concluded that a 3-layer neural network with 2 to 3 hidden layer neurons for each input neuron was sufficient to produce reasonable and acceptable results. Initially we tried with 2 times the number of hidden layer neurons as a 15-30-1 neural network architecture. The results were excellent and closely followed the monitored pattern of real time data. We increased the number of neurons in the hidden layer and the results produced at some periods within the simulations were overfitted. Hence, we decided to use the 15-30-1 architecture in the remaining analysis.

5 ANN analysis

We considered four scenarios to analyze the impact of pumping on the salinity at Carquinez Strait. The three prediction data sets for this experiment were from 1997 and the validation data set (Validation data set: A set of examples used only to assess the performance [generalization] of a fully-specified neural network.) was from 1998. Two data sets represent spring and the remaining two data sets represent middle of summer. The data sets represent the continuous time series. To study the impact of potential management options on salinity at Carquinez strait, the pumping rate at the federal and state projects was increased by 50% for all four periods. The salinity at Carquinez Strait was predicted to increase from 3-20%.

When the pumping is increased by 50% for the two input variables out of 15 input variables considered in training, occasionally the prediction of neural network results are under fitted. Under fitting is a result of data for some period that exceeds the range of values encountered during the neural network training (for example April 1997). Overall there is a possibility of increase of salinity up to 20% at Carquinez strait. The prediction for the above analysis is shown in (Figure 2).

6 Discussion

After experimenting with different scenarios, a three layer neural network with two hidden layer neurons for each input neuron with Levenberg-Marquardt algorithm was determined for this problem to be the best.

Nicky et.al. [12] used daily data for the period 1971-1991 in recurrent feed forward neural network to model salinity intrusion. Any missing values in the data were filled with linearly interpolated data from the nearest available data values. When considering for the San Francisco bay delta, daily data were inappropriate. The two high and two low tides each day in the San Francisco bay delta system effects in the bay cannot be addressed clearly. The selection of one-hour data intervals will be able to simulate the problem much more effectively than daily intervals. One-hour data interval prediction of salinity will help the management authorities to regulate the pumping at state and federal pumping stations more efficiently, because pumping can be adjusted much more frequently and salinity can be managed much more efficiently.

The data available for years prior to 1996 are of insufficient quality to test for long-term trends in salinity in the bay-delta region. Data are collected with different sampling frequencies. The data collection from different agencies should have a common time interval of collection throughout the year. Additionally, there are significant removals of water from the delta that are not monitored that could play a role in determining the salinity of the bay-delta system. Adequate predictions of salinity may depend on obtaining these data. However, overall the predictions of the neural network model were quite good and could be used as a real time method to explore different management options for the bay-delta system.

Acknowledgement

Funding was provided by a grant from the U.S. EPA (EPA-R-825285-01-0).

Appendix I

Locations in Bay-Delta Region

RSAC045 Selby (Wickland Oil Pier)
SHWSF001 Central Bay at Presidio Fort Point
RSAC054 Sacramento River at Martinez (Carquinez strait)
RSAC064 Sacramento River at Port Chicago
SLMZU011 Montezuma Slough at Beldons
RCSM075 Consumnes River at Michigan Bar
RSAC101 Sacramento River at Rio Vista Bridge
RSAC155 Sacramento River at Freeport
CHGRL009 Grantline Canal at Tracy Blvd Bridge
ROLD040 Old River at Clifton Court Ferry
RSAN007 San Joaquin River at Antioch betweens lights 7 & 8
RSAN112 San Joaquin River at Vernallis
CHDMC006 Delta Mendota Canal at the head (beginning) of the concrete liner
CHSWP003 State water project California at Harvey O.Banks Delta Pumping Plant
CHDMC004 CDEC TRD Delta Mendota Canal at Tracy Pumping plant
CHCCC006 CCWD Pumping station, Contra Costa Canal at Pumping Plant 1

References

1. California Department of Water Resources. DWRDSM. http://wwwdelmod.water.ca.gov.
2. Francis Chung. Modeling Flow-Salinity relationships in the Sacramento-San Joaquin delta using artificial neural networks. Technical Report OSP-99-1. Department of Water resources, Office of SWP Planning, 1999.
3. Freeman J. A. and Skapura D. M, Neural networks and algorithms - Applications and Programming techniques. Addison-Wesley Publishing Company, 1992.
4. Hagan M. T., Demuth H.B. and Beale M. H, Neural Network Design. Boston, MA. PWS Publishing, USA, 1996.
5. Interagency ecological website. http://www.iep.ca.gov.
6. Jondarr. C, Back propagation Family. Technical Report C/TR96-05. Macquarrie University, 1996.
7. Kartalopoulos S.V, Understanding neural networks and fuzzy logic - basic concepts and applications. IEEE Press, USA, 1996.
8. Lau C, Neural Networks theoretical foundations and analysis. IEEE Press, 1992.
9. MacKay D. J. C, Bayesian Interpolation. Neural Computation Vol 4 No 3 415-447, 1992.
10. Martin T. Hagan and Mohammad B. Menhaj, Training Feed forward networks with the Marquardt Algorithm. IEEE Transactions on neural networks, Vol 5 No 6 989-993, 1994.
11. Master T, Practical neural network recipes in C++. Academic Press Inc. 1993.
12. Nicky Sandhu and Ralph Finch, Application of artificial neural networks to the Sacramento-San Joaquin Delta. Proceedings of fourth international conference on estuarine and coastal modeling. Oct 26-28, Sandi ego, California, USA 490-504, 1995.
13. Rumelhart D.E., G.E.Hinton and R.J. Williams, "Learning internal representations by error propagation", in D.E. Rumelhart and J.L. McClelland, Eds. Parallel Data Processing. Vol 1, Cambridge, MA: The MIT Press, 318-362, 1986.
14. Shumway, Robert H. and Rahman S. Azari, Statistical Analysis of historical delta water quality data. Report prepared for delta modeling section, Division of planning, California Department of water resources under contract No. B-59259, 1993.
15. Winkler, Edward D, Preliminary development of a statistical approach to salinity modeling in the western Sacramento-San Joaquin Delta and Suisun Bay California department of water resources Technical information record study code no 1463-CD-04, 1985.

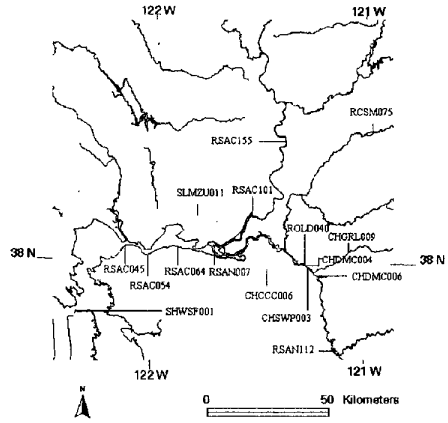

Figure 1. San Francisco Bay-Delta Region

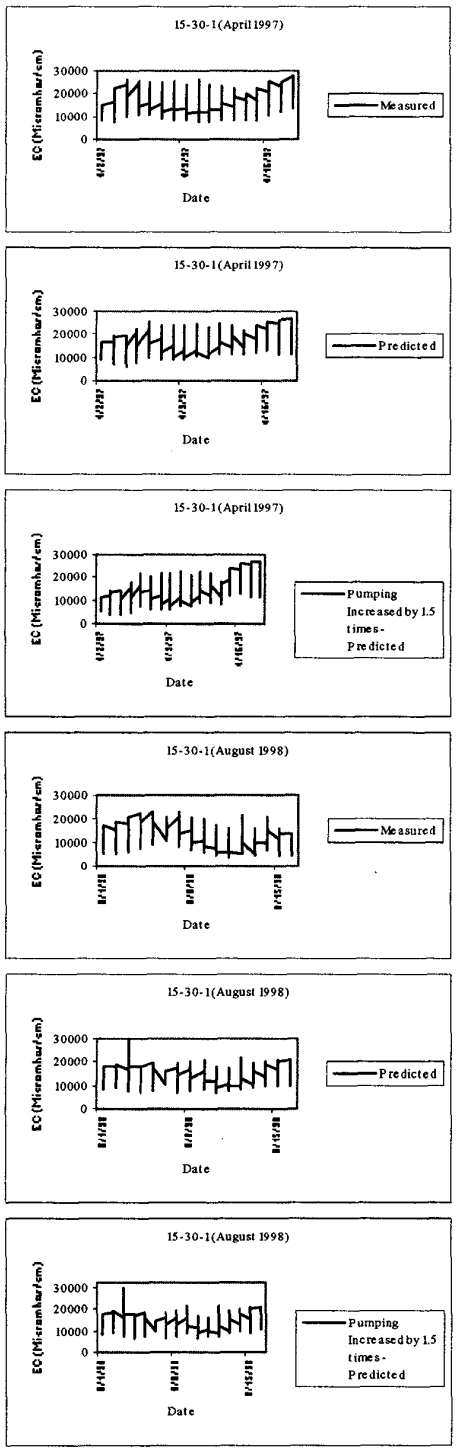

Figure 2 Real time measured and Predicted salinity at Carquinez strait using neural network for the period of April 1997 and August 1998

[1] Work done during Post-doc at UC Davis