Predicting Salinity in the Chesapeake Bay Using Neural Networks

by

Bruce L. Golden
Smith School of Business
University of Maryland

10/93
Goal

Construct multiple regression models and neural network models that accurately describe the dynamics of salinity in the Maryland portion of the Chesapeake Bay

- Other efforts use time series methods to predict surface and bottom salinity as part of a Bay water quality model
Source of Data

• Data collected by USEPA in five “regions” of the Chesapeake Bay
  Upper, middle, lower tributaries, and entire Bay
  18 stations in the mainstem Bay
  16 stations in tributaries
Source of Data -- continued

- Water samples collected at the bottom of the Bay (bottom data) and at various depths in the Bay (total data)
  - Old data: 36,000 observations 1984-1989
  - New data: 7,000 observations 1989-1990
Source of Data -- continued

- Ten different regression models and ten different neural network models are built using the old data
  5 regions x 2 depths
- Neural network models and regression models are compared using 20 data sets (old data and new data)
Figure 1. Survey stations in the Chesapeake Bay
Regression Models

• Extensive screening phase for independent variables

Four key independent variables

Day: day of the year on which measurements were taken
Depth: depth at which measurements were taken
Latitude: latitude of sampling station
Longitude: longitude of sampling station
Regression Models -- continued

• Used stepwise regression in SPSS/PC

  Avoid highly correlated independent variables

  Keep models simple: don’t include variables that add little in predictive power
Regression Models

- Constructed 5 bottom-data models and 5 total-data models using old data

- Entire Bay model using 36,000 observations

\[ R^2 = 0.649 \]
\[ \text{Salinity} = 199.839 - 1.151\text{Day1} + 1.161\text{Day2} + 0.283\text{Depth} - 4.863\text{Latitude} - 1.543\text{Longitude} - 13.402\text{Longitude1} \]
Regression Models -- continued

• Six independent variables in each model
  All coefficients were significant
  Each model easily passed an F test
  No problems with multicollinearity
  $R^2$ values ranged from 0.56 to 0.81
Neural Network Models

• Neural network configuration

<table>
<thead>
<tr>
<th>Station Number</th>
<th>Depth</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Date</th>
<th>Longitude x Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Salinity Level

Hidden Nodes
Neural Network Models --continued

• Neural network details
  
  Multilayer feedforward network
  Training by backpropagation
  Length of training session – 2000 iterations
  Training time on Sun 4/370 – 5 minutes
  Input value mapped to [-1, +1]
  Output (salinity) values mapped to [0, +1],
  same range as sigmoid function
Neural Network Models

- Neural Network parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Upper</th>
<th>Middle</th>
<th>Lower</th>
<th>Tributaries</th>
<th>Entire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning rate</td>
<td>.80</td>
<td>.60</td>
<td>.60</td>
<td>.20</td>
<td>.80</td>
</tr>
<tr>
<td>Momentum term</td>
<td>.40</td>
<td>.70</td>
<td>.10</td>
<td>.10</td>
<td>.10</td>
</tr>
<tr>
<td>Hidden nodes</td>
<td>.40</td>
<td>.30</td>
<td>.50</td>
<td>.30</td>
<td>.40</td>
</tr>
<tr>
<td>Slope</td>
<td>.80</td>
<td>.80</td>
<td>.80</td>
<td>.80</td>
<td>.80</td>
</tr>
</tbody>
</table>
Neural Network Models -- continued

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Upper</th>
<th>Middle</th>
<th>Lower</th>
<th>Tributaries</th>
<th>Entire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning rate</td>
<td>.20</td>
<td>.80</td>
<td>.80</td>
<td>.60</td>
<td>.20</td>
</tr>
<tr>
<td>Momentum term</td>
<td>.10</td>
<td>.80</td>
<td>.40</td>
<td>.20</td>
<td>.10</td>
</tr>
<tr>
<td>Hidden nodes</td>
<td>.30</td>
<td>.40</td>
<td>.40</td>
<td>.20</td>
<td>.30</td>
</tr>
<tr>
<td>Slope</td>
<td>.80</td>
<td>.80</td>
<td>.80</td>
<td>.80</td>
<td>.80</td>
</tr>
</tbody>
</table>
Neural Network Models -- continued

- Training the neural network

<table>
<thead>
<tr>
<th>Region of the Bay</th>
<th>Upper</th>
<th>Middle</th>
<th>Lower</th>
<th>Tributaries</th>
<th>Entire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% in training set</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td># in training set</td>
<td>199</td>
<td>243</td>
<td>190</td>
<td>79</td>
<td>271</td>
</tr>
</tbody>
</table>

| Total Data        |       |        |       |             |        |
| % in training set | 2     | 2      | 2     | 2           | 1      |
| # in training set | 250   | 330    | 280   | 78          | 363    |
Comparison of Models

• Regression models can use a different set of six independent variables in each region

• Neural network models are based on the same set of six variables in each region

• Computational results

<table>
<thead>
<tr>
<th></th>
<th>Range of Average Percent Absolute Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10 old data sets</td>
</tr>
<tr>
<td>Neural Network</td>
<td>9.54 – 16.18</td>
</tr>
</tbody>
</table>
Comparison of Models -- continued

- Key points
  - Neural network models have lower average PAE than the regression models in 18 out of 20 cases
  - Worst errors of the neural network models are not as bad as those from regression
  - Neural network models yield more errors in the 0-10% range than regression models
Conclusions

• Current combinations of training parameters work quite well for the neural network models

• Major advantage of the regression models is that they are easily explained

• Based on a small number of observations and six fixed variables, the neural network models predict salinity levels more accurately than do the regression models