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)

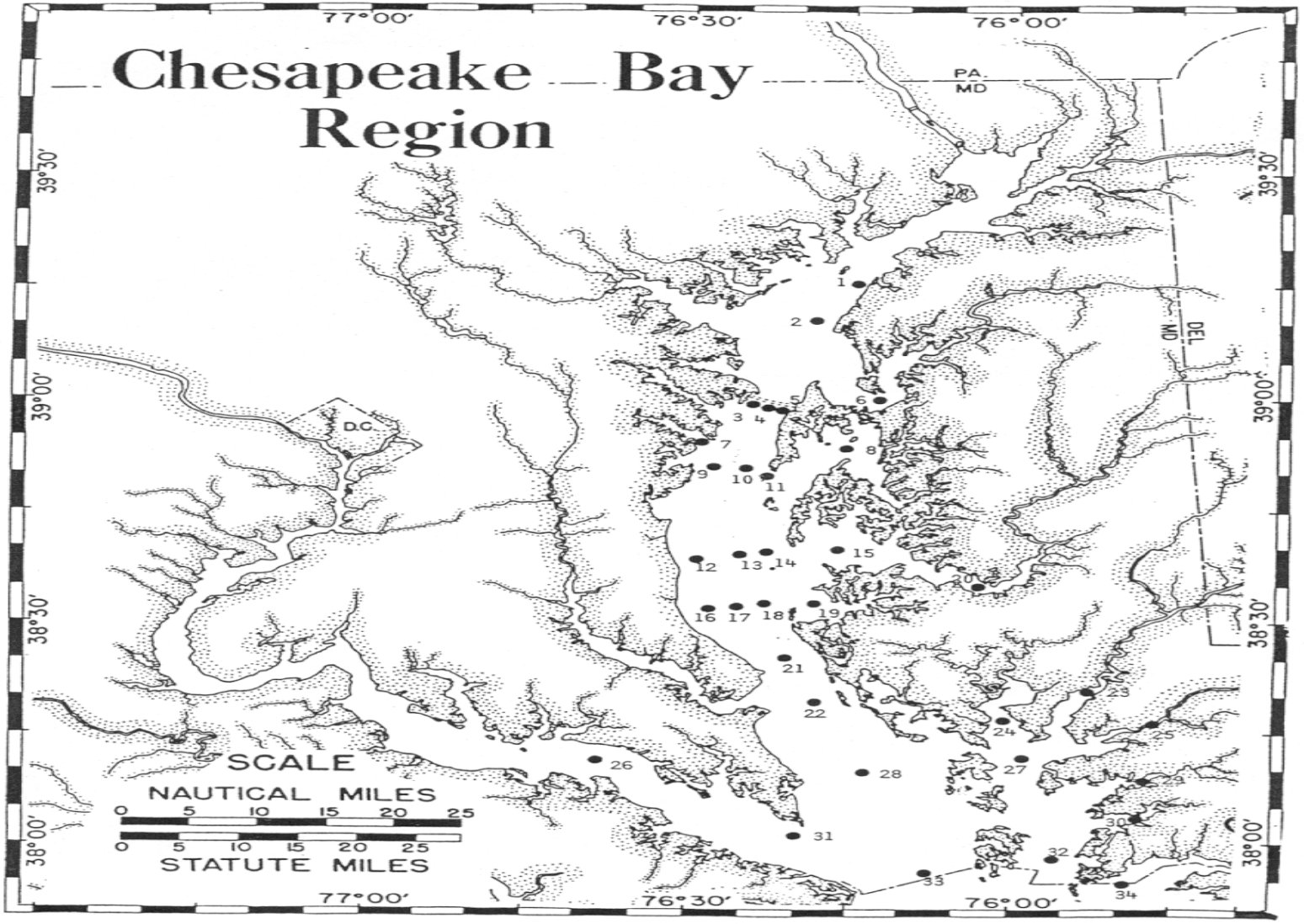


Figure 1. Survey stations in the Chesapeake Bay

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

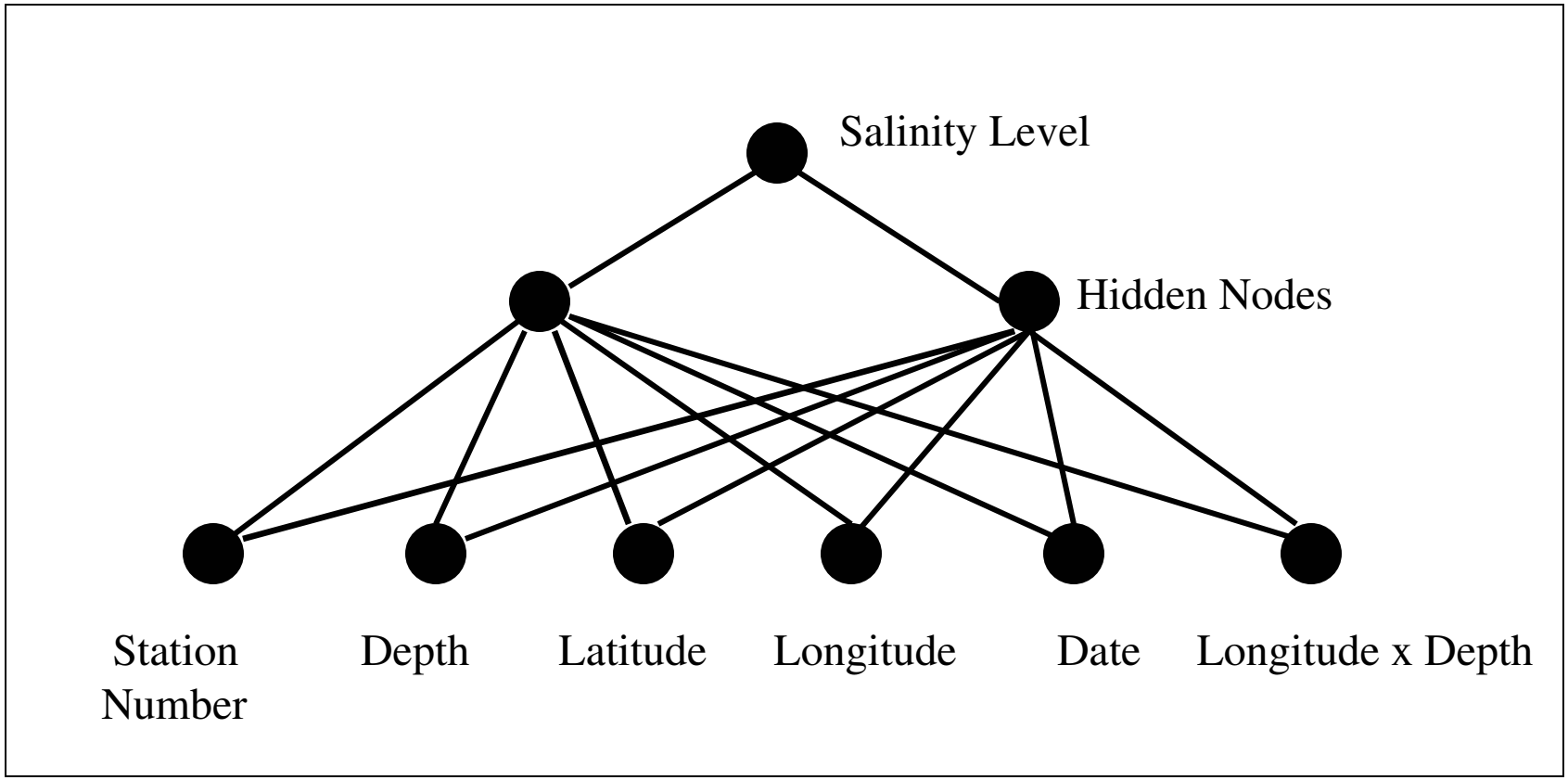
  Each model easily passed an F test

  No problems with multicollinearity

  R^2 values ranged from 0.56 to 0.81

Neural Network Models

- Neural network configuration



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

Bottom Data	Region of the Bay				
Parameter	Upper	Middle	Lower	Tributaries	Entire
Learning rate	.80	.60	.60	.20	.80
Momentum term	.40	.70	.10	.10	.10
Hidden nodes	.40	.30	.50	.30	.40
Slope	.80	.80	.80	.80	.80

Neural Network Models -- continued

Total Data	Region of the Bay				
Parameter	Upper	Middle	Lower	Tributaries	Entire
Learning rate	.20	.80	.80	.60	.20
Momentum term	.10	.80	.40	.20	.10
Hidden nodes	.30	.40	.40	.20	.30
Slope	.80	.80	.80	.80	.80

Neural Network Models -- continued

- Training the neural network

	Region of the Bay				
	Upper	Middle	Lower	Tributaries	Entire
Bottom Data					
% in training set	20	20	20	20	10
# in training set	199	243	190	79	271
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

	Range of Average Percent Absolute Errors	
	10 old data sets	10 new data sets
Regression	9.60 – 16.46	9.19 – 20.15
Neural Network	9.54 – 16.18	7.70 – 19.37

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