THÜNEN
Digitalization sponsored by Thünen-Institut

THIS PAPER IS NOT TO BE CITED WITHOUT PRIOR REFERENCE TO THE AUTHORS

INTERNATIONAL COUNCIL FOR
THE EXPLORATION OF THE SEA

ANADROMOUS AND CATADROMOUS
FISH COMMITTEE
CM 1991/M: 10

THE USE OF A NEURAL NETWORK TO DISTINGUISH NORTH AMERICAN AND EUROPEAN SALMON (*SALMO SALAR L.*) USING SCALE CHARACTERISTICS

E C E Potter[1], L Kell[1] and D G Reddin[2]

[1]Ministry of Agriculture, Fisheries and Food
Fisheries Laboratory,
Pakefield Road, LOWESTOFT, Suffolk
United Kingdon

[2]Science Branch
Department of Fisheries and Oceans
PO Box 5667
St. John's Newfoundland A1C 5X1, Canada

Bundesforschungsanstalt
Bibliothek
für Fischerei, Hamburg

ABSTRACT

Electrophoretic separation of protein polymorphisms and discriminant analysis of scale characters are both routinely used to discriminate the continent of origin of Atlantic salmon caught at West Greenland. However, the use of electrophoretic separation to discriminate fish in the commercial catches is precluded by the difficulty of collecting sufficient samples. Sampling for the discriminant analysis is simpler, although this method provides only about an 80% correct classification for both North American and European fish. This paper presents an alternative method for analysing the scale data, based on the use of a neural network. The same data sets were used to develop and test a discriminant function and a neural network, and the results were compared. The network was set up with the four input variables (river age, fork length and two scale circuli counts) grouped into discrete classes in order to provide 33 input neurons; the 2 output neurons corresponded to 'North American' and 'European'; and the network was given 17 hidden layer neurons. The neural network analysis gave more accurate separation (85.8% correct) of all samples into North American and European groups than the discriminant analysis (80.3% correct). The classification by the network was improved by reducing the testing tolerance level and thus removing the most doubtful results from the assessment. At a tolerance level of 0.2, 15% of the samples were unclassified, but the accuracy of the remaining classifications was improved to almost 90%.

INTRODUCTION

The West Greenland fishery for Atlantic salmon (Salmo salar L.) exploits fish originating from both North American and European rivers. This presents a complex management problem and has led to the requirement to find methods to distinguish fish from the two continents in order to estimate the composition of the catch. Both protein polymorphisms and scale characters have been used extensively to discriminate between stocks or groups of stocks of Atlantic salmon (eg Payne and Cross, 1977; Lear and Sandeman, 1980; Reddin et al, 1988)). Electrophoretic separation of protein polymorphisms has been shown to provide fairly reliable discrimination between North American and European fish (Verspoor and Reddin, 1989). About 65% of each stock group can be assigned with a

probability of >1000:1 of being correct and a further 30% with a probability of >10:1 of being correct. However, the routine use of this method in the West Greenland fishery is precluded because it is impossible to obtain the necessary tissue samples from a large enough number of fish. As a result, discriminant analysis of scale characteristics in currently used in the annual assessment of the fishery (Reddin, 1986; Anon, 1989). The discriminant function is developed and tested using data from salmon whose origin has been established by electrophoretic separation (Reddin et al, 1988) or tag recoveries (eg Russell et al, 1990).

The purpose of this paper is to compare the accuracy of the discriminant analysis with an alternative method, that of 'neural networks', in the analysis of the same data set.

NEURAL NETWORKS

Neural networks are computer models which are designed to operate in a similar way to the brain by modelling mathematically the functions of neurons and synapses. Neurons in the brain are highly interconnected, and it appears to be this property that gives them their ability to learn and recognise patterns.

The basic mode of operation of biological neurons is that they accept and combine many inputs; if sufficient inputs are received at the same time, then the neuron will produce an output, otherwise the neuron will remain inactive. The efficiency of the contact between each input and the neuron is varied by chemical changes where they join at the synapses. The variable efficiency of biological synapses can be mimicked mathematically in the artificial neuron by applying a multiplicative weighting factor to each of the inputs; an 'efficient synapse' gets a high weighting, while a 'weak synapse' gets a low weighting. Thus:

$$\text{Total input} = \sum_{i=1}^{n} w_i I_i$$

where w_i is the weighting value on the ith input (I_i).

This sum is then compared with a threshold value, θ, for the neuron; if this value is exceeded the output will be 1, if not it will be 0.

Such an artificial neuron can be trained by presenting it with sets of input and output patterns (or vectors):

e.g. Input pattern:

Length	65
River Age	3
Scale parameter 1	23
Scale parameter 2	13

Output pattern:

American	1
European	0

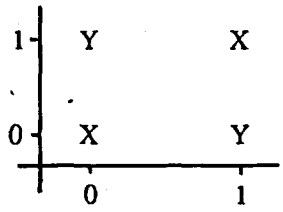
The neuron is initially set up with random weights (w_i) on the inputs. During the training phase, the input pattern is used to produce an output, which is then compared with the correct output. If the difference between the network output and the target is within acceptable limits, then no 'learning' takes place. If, however, the difference exceeds the limit, then the weighting factors are adjusted to reduce the difference.

This learning algorithm allows the artificial neuron to divide the pattern space in a simple manner; for example it can separate patterns in two dimensions with a line. This is much the same as a discriminant function. However, there are limitations to this type of discrimination. One insoluble problem is the exclusive-OR (XOR) function,

which produces one output (Y) if only one of its two inputs is 'on' but another (X) if both are 'on' or both are 'off'. The input and output vectors are shown with a graphical representation below:

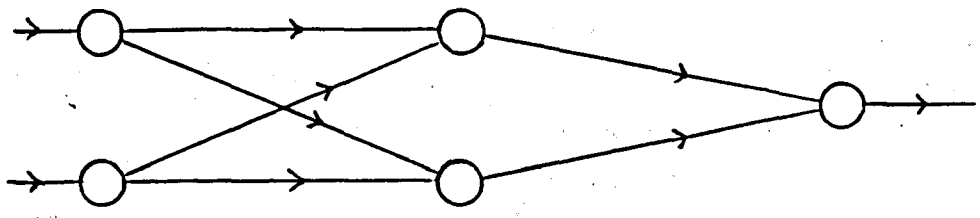
	Input patterns:	Output patterns:
1.	0, 0	X
2.	0, 1	Y
3.	1, 0	Y
4.	1, 1	X

Graphical representation:



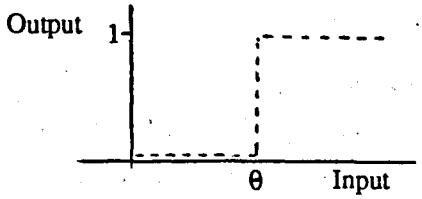
In the graph, no single line can be drawn to separate the 'X' values from the 'Y' values. For the single neuron, there are no possible weighting values for the inputs that will produce the same output pattern when inputs are on or off; thus it cannot be trained to discriminate these groups.

The constraint of linear separability was overcome by the development of multilayer neural networks. In these models, neurons (indicated by circles in the diagram below) are arranged in, usually, three layers; an input layer, a middle 'hidden' layer and an output layer.

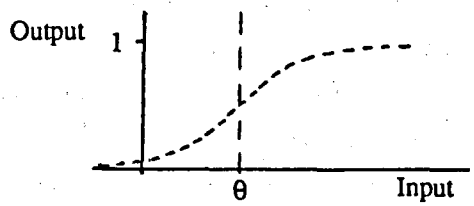


Input layer Hidden layer Output layer

Each unit in these layers is like the single artificial neuron described above, although the threshold function has to be changed. If, for example, the neurons in the hidden layer only gave outputs of '1' or '0', the input layer would be masked from the output layer, and no information would be passed to show which neuron weightings had to be adjusted. The figures below show the step threshold function and an alternative non-linear function of the type required for multilayer networks:



Step threshold function



Non-linear threshold function.

Increasing the number of layers increases the resolving power of the network. As explained above, a single neuron can define a single line in the pattern space (visualising the problem in two dimensions). In a two layer

system several neurons (each defining a line) are combined, producing more complex space partitioning. Regions defined in this way are called 'convex hulls' and have the property that any two points within them can be joined by a straight line that does not cross the boundary of the region. If a third layer of neurons is added, the space defined will be a combination of 'convex hulls', which can produce arbitrary complex shapes in the pattern space. This means that adding further layers will not help improve the resolving power of the network.

METHODS

The database:

Both discriminant analysis and the neural network analysis are based on samples of 'known' origin. The database used in this comparison consisted of samples from North American and European origin 2 sea-winter salmon collected in homewater fisheries (1980-85) and 1 sea-winter salmon sampled at West Greenland (1986-89). Data were provided on sample year, fork length, river age, sea age and two scale variables. Fork length values for fish sampled in homewaters were not appropriate for the analysis and were therefore given as 'unknown'. Fish sampled at Greenland were identified to continent of origin by electrophoretic separation or tag recoveries. Scales were taken from the same location on the fish, and circuli counted as described by Reddin (1986). The scale variables used were the circuli counts in the winter growth portion (CS1W) and in the summer growth portion (CS1S) of the first sea year, read at 45° from the longest axis.

Because the discriminant analysis was to be carried out on separate river age classes (see below) the data set was established with equal numbers of North American and European salmon of river age 1 (200), 2 (330) and 3-6 (220); this gave a total of 750 salmon from each continent. The database was divided randomly in half to provide a 'training' set and a 'test' set. In both the discriminant and neural network analyses, the roles of these sets were reversed, and the training and testing processes were repeated; the results were then combined, and misclassification and error rates were based on a summation of both runs (Pella and Robertson, 1979).

The misclassification rate is defined as the sum of the incorrectly classed samples divided by the total number of samples classified. The error rate is the actual proportion of North American or European salmon in the samples minus the calculated proportions classified.

Discriminant analysis:

The parametric discriminant analysis model employed was as described by Reddin (1986). Because the results of an analysis of variance (ANOVA) indicate significant effects of river age on the scale characters CS1S and CS1W, three discriminant functions were developed for salmon of river age 1, 2 and 3-6 respectively.

The discriminant model used was based on a measure of generalised squared distance (Rao, 1973) employing either the individual within-group covariance matrices or the pooled covariance matrix. For river age 1 salmon the within-covariance matrix was used for classification because a test of the homogeneity of the within-covariance matrices was significant at the 10% level ($X = 27.2$, $df = 3$, $p < 0.01$) (Kendall and Stuart, 1961). For older river age salmon, pooled covariance matrices were used because tests of the homogeneity of the within-covariance matrices were not significant at the 10% level (for river age 2 salmon, $X = 2.06$, $df = 3$, $p = 0.15$, and for river age 3-6 salmon, $X = 0.14$, $df = 3$, $p = 0.71$).

Neural network analysis:

Neural networks operate better on discrete binary inputs rather than continuous variables, and the learning power is improved if the number of inputs is increased. For this reason each of the input variables was divided into the following classes:

CS1S	<4	CS1W	<18	River age 1	Fork length <57
CS1S	4	CS1W	18-19	River age 2	Fork length 58-59
CS1S	5	CS1W	20-21	River age 3	Fork length 60-61
CS1S	6	CS1W	22-23	River age 4	Fork length 62-63
CS1S	7	CS1W	24-25	River age >4	Fork length 64-65
CS1S	8	CS1W	26-27		Fork length 66-67
CS1S	9	CS1W	28-29		Fork length 68-69
CS1S	10	CS1W	30-31		Fork length >69
CS1S	11	CS1W	>31		Fork length Unknown
CS1S	>11				

For each record (salmon), these classes were given values of '1' if the fish measurement was within the given range or '0' if it was not. Each of the groups was assigned to an input neuron, giving a total of 33 neurons in the input layer.

The hidden and output layers had 17 and 2 neurons respectively. The two output neurons corresponded to 'American' and 'European'; the 'correct' outputs were defined such that, if the fish was American, the American output neuron would be given a value '1' and the European output neuron a value '0'. If the fish was European the reverse would be true.

Two networks were trained by repeatedly presenting the examples in the respective 'training' data sets. The learning tolerance (ie the accuracy required for an example to be considered correctly learnt) was set at 0.1. The trained networks (effectively the values of w_i) were stored at regular interval during the training procedure and tested on the alternate 'test' data sets. The tolerance during testing was set at a lower level than during learning, for example 0.4; at this level, if the network assigned a value of over 0.6 to the American output (and less than 0.4 to the European output) it would identify the fish as American, and vice versa.

RESULTS

Discriminant analysis:

The results of classifying the 'test' samples for the three river age groups are shown below, and the total classification matrix for all age groups combined is given in Table 1:

River age	Misclassification rate	Error rate
1	10.0%	+/- 3.5%.
2	23.3%	+/- 3.0%
3-6	23.2%	+/- 0.9%
All age groups	19.7%	+/- 2.0%

Neural network:

Figure 1 shows the change in the number of samples correctly learnt on successive training runs through the two 'training' data sets; the number increases quickly at first and then more slowly, levelling off after about 100 runs. The change in the proportion of the 'test' data sets that were correctly identified by the two networks after different numbers of training runs is shown in Figure 2. The number of records correctly identified improves very rapidly for the first 5 training runs, followed by a slow improvement for about another 35 runs; with further training the discriminatory power of the networks gradually deteriorates.

From Figure 2, the two networks were judged to be optimally trained after 40 runs of the 'training' data sets. Tables 2-4 show the results of testing these networks on their respective 'test' data sets at different tolerance levels. In Table 2, each record is classified according to the output neuron (American or European) with the highest value. In Tables 3 and 4, classification is made using tolerance levels of 0.4 and 0.2 respectively, and there are therefore some unclassified records. The misclassification and error rates from the tests are summarised below:

Tolerance level	Misclassification rate	Error rate	Unclassified
None	14.2%	+/- 0.2%	0%
0.4	12.7%	+/- 0.4%	4.2%
0.2	10.1%	+/- 1.3%	14.9%

DISCUSSION

The neural network analysis gave more accurate separation (85.8% correct) of all samples into North American and European groups than the discriminant analysis (80.3% correct). This is as predicted because the network is able to separate the pattern space in a more complex manner than the discriminant analysis. The classification by the network was improved by reducing the testing tolerance level and thus removing the most doubtful results from the assessment. At a tolerance level of 0.2, 15% of the samples were unclassified, but the accuracy of the remaining classifications was improved to almost 90%. This gives a misclassification rate equal to about half that of the discriminant analysis. The misclassification rate by the neural network could probably be reduced still further by using an even lower tolerance level, but the possibility that the unclassified samples may be biased towards particular stocks needs to be investigated.

It should be noted that, although length measurements were not available for about half of the samples, the neural network operated successfully on the incomplete data sets. This demonstrates the flexibility of this technique and illustrates how a wide range of quantitative and descriptive data could be used in a neural network analysis.

At present, the only way to fine-tune the network in order to improve its discriminatory powers is by trial and error. For example, increasing the number of input neurons by putting the data into classes improves the learning power of the network, but it also reduces the number of examples of each input pattern for the network to train on. A compromise also has to be made in the number of training runs the network is given. The 'training curve' in Figure 1 shows that the network may get progressively better at discriminating the samples in the 'training' data set over many training runs, although there may always be some records that cannot be learnt, if, for example, an American and European fish have almost identical input patterns. The learning curve in Figure 2, however, shows that the

discriminatory power of the network may only over about 40 learning runs, but also demonstrates that the network can be overtrained. It appears that by learning the differences in the 'training' data set too precisely, the network looses its ability to generalise on other similar data sets.

The result of this comparative study suggests that neural networks may offer a useful alternative to discriminant analysis in other situations, for example, in conventional scale reading. Modern image analysis systems permit the objective collection of many scale measurements with little effort. Such data would be ideal for analysis by neural networks which learn well with many inputs even when the data are imprecise.

REFERENCES

ANON, 1990. Report of the North Atlantic Salmon Working Group. ICES C.M. 1990/Assess:11

KENDALL, M.G. AND STUART, A. 1961. The advanced theory of statistics, Vol. 3. Charles Griffin and Co., London.

LEAR, W.H. AND SANDEMAN, E.J. 1980. Use of scale characters and discriminant functions for identifying continental origin of Atlantic salmon. Rapp. P.-v. Reun. Cons. int. Explor. Mer., 176: 68-75.

PAYNE, R.H. AND CROSS, T.F. 1977. Liver asparate aminotransferase polymorphism: a new tool for estimating proportions of European and North American salmon at West Greenland. ICES C.M. 1977/M:10

PELLA, J.J. AND ROBERTSON, T.L. 1979. Assessment of composition of stock mixtures. Fish. Bull. U.S., 77: 387-398.

RAO, C.R. 1973. Linear statistical inference and its applications. John Wiley and Sons, New York. 522pp.

REDDIN, D.G. 1986. Discrimination between Atlantic salmon (*Salmo salar* L.) of North American and European origin. J. Cons. int. Explor. Mer., 43: 50-58.

REDDIN, D.G., STANSBURY, D.E. AND SHORT, P. 1988. Continent of origin of Atlantic salmon (*Salmo salar* L.) at West Greenland. J. Cons. int. Explor. Mer., 44: 180-188.

RUSSELL, I.C., POTTER, E.C.E., REDDIN, D.G. AND FRIEDLAND, K.D. 1990. Recoveries of coded wire microtags from salmon caught at West Greenland in 1989. ICES C.M.1990/M:20

VERSPOOR, E. AND REDDIN, D.G. 1989. A model for the classification of Atlantic salmon *Salmo salar* to continent of origin using genetic protein variation. ICES Working Group on North Atlantic Salmon Working Document 1989/33

Table 1. The results of classifying test samples of Atlantic salmon to continent of origin by discriminant analysis of scale characteristics; results combined for river age 1, river age 2 and river age 3-6 fish.

Actual	Predicted origin			
Origin	American	European	Unclassified	Total
American	617 (82.3%)	133 (17.7%)	0	750
European	163 (21.7%)	587 (78.3%)	0	750
Totals classified	654 (52.0%)	622 (48.0%)		
Actual	750 (50.0%)	750 (50.0%)		
Error rate	+2.0%	-2.0%		

Number of records classified: 1500 (100%)

Number of records correctly classified: 1204 (80.3%)

Number of records incorrectly classified: 296 (19.7%)

Table 2. The results of classifying test samples of Atlantic salmon to continent of origin using a neural network with no threshold level; test fish assigned to the origin given the highest output value.

Actual	Predicted origin			
Origin	American	European	Unclassified	Total
American	645 (86.0%)	105 (14.0%)	0	750
European	108 (14.48%)	642 (85.6%)	0	750
Totals classified	753 (50.2%)	747 (49.8%)		
Actual	750 (50.0%)	750 (50.0%)		
Error rate	+0.2%	-0.2%		

Number of records classified: 1500 (100%)

Number of records correctly classified: 1287 (85.8%)

Number of records incorrectly classified: 213 (14.2%)

Table 3 The results of classifying test samples of Atlantic salmon to continent of origin using the neural network with a testing tolerance level of 0.4.

	Predicted origin			
Actual Origin	American	European	Unclassified	Total
American	631 (84.1%)	90 (12.0%)	29 (3.9%)	750
European	93 (12.4%)	623 (83.1%)	34 (4.5%)	750
Totals classified	724 (50.4%)	713 (49.6%)		
Actual	750 (50.0%)	750 (50.0%)		
Error rate	+0.4%	-0.4%		

Number of records classified: 1437 (95.8%)

Number of records correctly classified: 1254 (87.3%)

Number of records incorrectly classified: 183 (12.7%)

Table 4 The results of classifying test samples of Atlantic salmon to continent of origin using the neural network with a testing tolerance level of 0.2.

	Predicted origin			
Actual Origin	American	European	Unclassified	Total
American	587 (78.3%)	62 (8.3%)	101 (13.5%)	750
European	67 (8.9%)	560 (74.7%)	123 (16.4%)	750
Totals classified	654 (51.25%)	622 (48.75%)		
Actual	750 (50.0%)	750 (50.0%)		
Error rate	+1.25%	-1.25%		

Number of records classified: 1276 (85.1%)

Number of records correctly classified: 1147 (89.9%)

Number of records incorrectly classified: 129 (10.1%)

Figure 1. Number of samples correctly learnt against the number of training runs for the two training data sets.

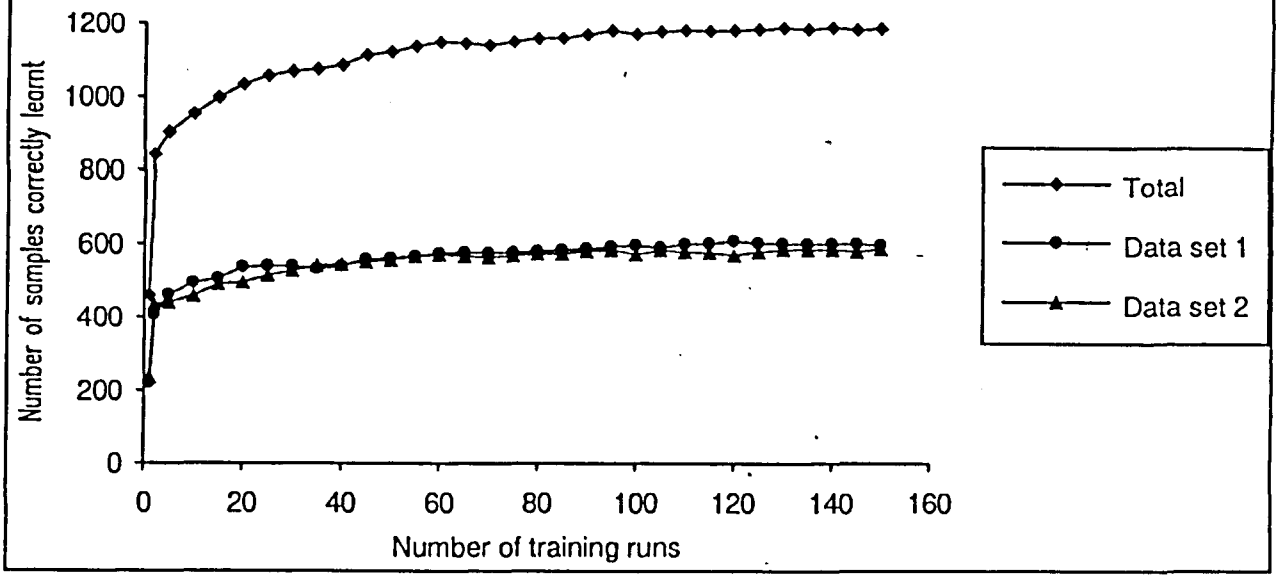


Figure 2. Number of samples correctly classified from two test data sets against number of training runs for the two training data sets.