25 Fish stock identification through neural network analysis of parasite fauna

Campbell, N., MacKenzie, K., Zuur, A.F., Ieno, E.N. and Smith, G.M.

25.1 Introduction

The main aim of fisheries science is to interpret relevant information on the biology of the species in question, records of fishing effort and size of catches, in order to predict the future size of the population under different fishing regimes, allowing fishery managers to make decisions on future fishing efforts. The common approaches to evaluation, modelling and management of fish stocks assume discrete populations for which birth and death are the significant factors in determining population size, and immigration and emigration are not (Haddon 2001). Consequently, for successful management of fisheries, it is vital that populations that conform to these assumptions are identified. In areas where two stocks mix, it is useful to be able to quantify this so that catches can be assigned to spawning populations in the correct proportions.

Several techniques have been used to identify discrete fish stocks and quantify their mixing, such as physical tags, microchemistry of hard parts and a range of genetic markers. See Cadrin et al. (2005) for a comprehensive review. There is no single "correct" approach to stock identification, the trend being towards multidisciplinary studies that apply a range of methods to the same set of fish to allow cross-validation of findings. One of the more popular methods involves the use of parasites as biological tags, and has been used for over 60 years (Herrington et al. 1939). This technique has several advantages over other methods, such as low cost, suitability for delicate species and straightforward sampling procedures. Its main disadvantage is the limited knowledge available on the life cycles and ecology of many marine parasites, but as research in these areas results in more and more information becoming available, the efficiency of the method increases accordingly (MacKenzie 1983; Lester 1990).

The theory behind this technique is that geographical variations in the conditions that a parasite needs to successfully complete its life cycle occur between areas and so between fish stocks (factors such as distribution of obligatory hosts in the life cycle, environmental conditions or host feeding behaviour). This leads to differences in parasite prevalence (the proportion of a host population infected with a particular parasite species), abundance (the average number of a particular

parasite species found per host) or intensity (the average number of parasites found in infected individuals) between areas.

A "classical" parasites-as-tags study involves carrying out a preliminary study to identify parasite species that vary in prevalence, abundance or intensity within the study area, followed by the collection of data from a larger number of fish over several years to produce conclusive evidence of a lack of mixing between different parts within the study area. Note that the absence of a difference in parasite prevalence, abundance or intensity between samples is not necessarily indicative of homogeneous mixing. This approach allows migrations, recruitment of juveniles or mixing of different spawning populations to be observed and quantified. The practical application of this method was modelled and verified by Mosquera et al. (2000). It is particularly useful in areas where a small but significant degree of mixing between two populations occurs, obscuring genetic differences between populations.

Several more recent studies have taken a more complex approach to the statistical treatment of parasites as tags of their host populations. These studies have considered each fish as a habitat and treated the entire parasite fauna in that individual as a community. Discriminant analysis (DA) is applied to the parasite abundance data of the community, in order to identify groups of similar fishes (Lester et al. 1985). Moore et al. (2003) had some success with the application of DA to the parasite fauna of narrow-barred Spanish mackerel (*Scomberomorus commersoni*) around the coast of Australia, in order to quantify movement and mixing of stocks.

Discriminant analysis, however, makes several assumptions that make it less suitable for this sort of analysis. First, for testing of hypotheses, DA assumes normality and homogeneity within each group of observations per variable (in this case, parasite abundances). Second, DA works best with roughly equal sample sizes and requires the number of variables to be less than the smallest sample size minus two.

25.2 Horse mackerel in the northeast Atlantic

The Atlantic horse mackerel (*Trachurus trachurus*) is a small pelagic species of fish, with a maximum size of about 40 cm. They are the most northerly distributed species of the jack-mackerels (family Carangidae) (FAO 2000), and they support a sizeable fishery in the northeast Atlantic, both for human consumption and for industrial processing. Catches in the region have been over 500,000 tonnes per year in recent times. They feed at a slightly higher trophic level than many small pelagic fishes, their diet consisting of planktonic copepods, small fishes and benthic invertebrates. This diverse diet is reflected in a diverse parasite community, and 68 taxa have been reported to infect *T. trachurus* (MacKenzie et al. 2004).

There has been uncertainty over the identity of stocks in the northeast Atlantic for over a decade. In this area, the International Council for the Exploration of the Seas (ICES) issues management advice for the horse mackerel, which assumes the

existence of three stocks. These are a (i) Western stock, (ii) a North Sea stock and (iii) a Southern stock (Figure 25.1).

These stocks have been defined mainly from observations of the distribution of eggs in regular plankton surveys, and on historical records of the distribution of catches (Eltink 1992; ICES 1992). Until recently, publications dealing with the definition of stock structure in horse mackerel were rare and covered only a small part of the species distribution. In the southern stock there are some works dealing with differences in anisakid infestation levels (Abaunza et al. 1995; Murta et al. 1995); whilst using allozymes, some authors found differences between areas in the northeast Atlantic (Nefedov et al. 1978); whereas others did not (Borges et al. 1993). A recent, EU-funded multidisciplinary study, HOMSIR, has resolved some of the problems with stock identity, but there are still unanswered questions.

One of the problems with stock definition for *T. trachurus* is that it is a highly migratory species, spawning along the edge of the continental shelf, in water around 200 m deep, then dispersing to feed over a wider area. It is thought that the Western and North Sea stocks overlap at certain seasons in the English Channel (Macer 1977), which may cause some degree of mixing between these stocks. Mixing between the Western and Southern stocks remains an unknown quantity, and there has been particular concern about the boundary between these stocks.

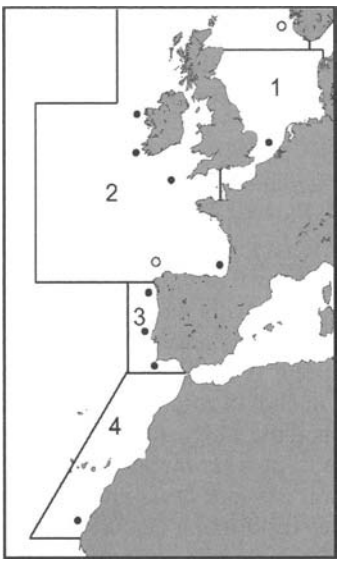

Figure 25.1. A graphical representation of the distribution of horse mackerel stock in the northeast Atlantic. 1. North Sea stock. 2. Western stock. 3. Southern stock. 4. African stocks, after ICES (1992, 2004). Locations of samples collected for verification of stock identity are marked with filled circles and those collected to investigate stock mixing with empty circles.

25.3 Neural networks

The original 'neural network' model was proposed in the 1940s (McCulloch and Pitts 1943), although it is only with the advent of cheap, powerful computers that this technique has begun to be applied widely. There has been a great deal of hype surrounding neural networks. They are, however, simply a non-linear statistical approach to classification. One of the attractions of neural networks to their users is that they promise to avoid the need to learn other more complex methods — in short, they promise to avoid the need for statistics! But this is a misconception: For example, extreme outliers should be removed, collinearity of variables should be investigated before training neural networks, and it would be foolish to ignore obvious features of the data distributions and summaries such as the mean or standard error. The neural network promise of 'easy statistics' is, however, partly true. Neural networks do not have implicit assumptions of linearity, normality or homogeneity, as many statistical methods do, and the sigmoid functions that they contain appear to be much more resistant to the effects of extreme values than regression based methods. Many of the claims made about neural networks are exaggerated, but they are proving to be a useful tool and have solved many problems where other methods have failed.

The name 'neural network' derives from the fact that it was initially conceived of as a model of the functioning of neurons in the brain — the components of the network represent nerve cells and the connections between them, synapses, with the output of the nerve switching from 0 to 1 when the synapses linking to it reach a 'threshold value'.

For the purposes of this chapter, a neural network can be thought of as a classification model of a real world system, which is constructed from the processing units ('neurons') and fitted by training a set of parameters, or weights, which describe a model that forms a mapping from a set of given values known as inputs to an associated set of values known as outputs (Saila 2005). The weights are trained by passing sets of input–output pairs through the model and adjusting the weights to minimize the error between the answer provided by the network and the true answer. A problem can occur if the number of training iterations, or 'epochs' is too large. This reduction in classification success of the data not used in training is known as over-fitting. Once the weights have been set by a suitable training procedure, the model is able to provide output predictions for inputs not included in the training set. The neural network takes all the input variables presented in the data and linearly combines them into a derived value, in a so-called 'hidden layer object' or node (Smith 1993). It then performs a nonlinear transformation of this derived value (Figure 25.2). The use of multiple hidden layer objects in a neural network allows different non-linear transforms of data, with each neuron (node) having its own linear combination, increasing the classifying power of the network.

Originally, the neuron was activated with a step function (represented as the dashed line in Figure 25.2), when the combined input values exceeded a certain value; however, this more flexible sigmoid function allows differentiation and

least squares fitting, leading to the back propagation algorithm, making it possible to tune the weights more finely.

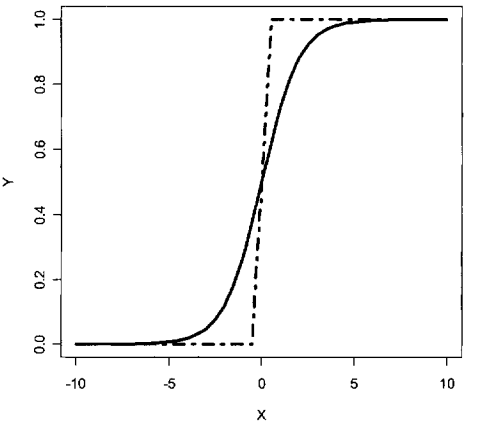

Figure 25.2. An example of the transformation that the hidden layer applies to linear combinations of the input data: $y = 1/(1 + \exp(-x))$. An example of the step function used in early neural network models is shown in the dashed line.

A possible example is shown in Figure 25.3. In this case, we are interested in knowing the stock composition of a mixed sample of fish, and we have count data on six species of parasites from these fish. These counts are treated as the six input variables to our network. The network has four units in the hidden layer. The neural network in such an example would need to have been trained with data from fish that we knew belonged to stocks X and Y beforehand if it was to work successfully. This type of neural network is referred to in the literature by many names, such as feed-forward network, multilayer perceptron, or simply vanilla neural network, named for the generic ice-cream flavour.

The number of units in a hidden layer is variable. Problems can arise if too few or too many units are used. If the network has too few units, it will not be flexible enough to correctly classify the data with which it is presented. On the other hand, if it has too many units, a problem known as over-parameterisation occurs; this reduces the chances of successful classification. Having many hidden layer objects also increases the computing power required to run the function. Often, a trial-and-error approach is used to determine the optimum number of hidden layer units for a particular dataset; however there are other methods that take a more considered approach, such as cross-validation, bootstrapping, and early stop.

Neural networks differ in philosophy from most statistical methods in several ways. A network typically has many more inputs than a typical regression model. Because these are so numerous, and because so many combinations of parameters result in similar predictions, the parameters can quickly become difficult to interpret and the network is most simply considered as a classifying 'black box'. This

means that areas where a neural network approach can be applied in ecology are widespread. They are less useful when used to investigate or explain the physical process that generated the data in the first place. In general, the difficulty in interpreting what the functions contained within these networks mean has limited their usefulness in fields like medicine, where the interpretation of the model is vital.

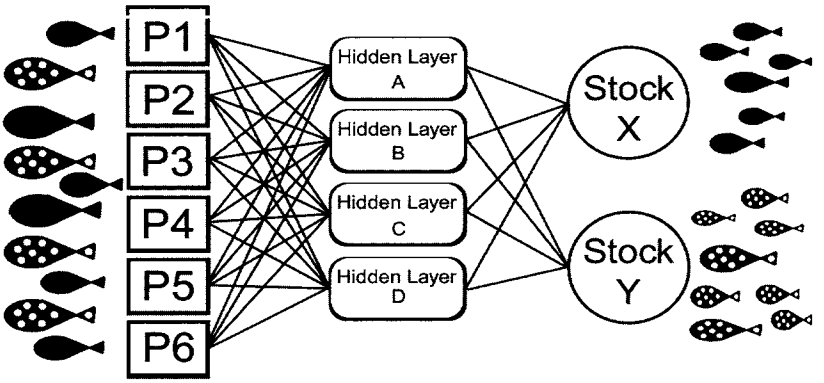

Figure 25.3. Graphical representation of the structure of a neural network. This one has six inputs (P1 to P6), one hidden layer with four neurons or units (A to D) and two output neurons into which it will classify the data (stocks *X* and *Y*).

There have been several uses of neural networks in fisheries science (Huse and Grøjsæter 1999; Maravelias et al. 2003; Engelhaard and Heino 2004), however, these have mainly attempted to predict changes in fish abundance, recruitment or distribution, based on environmental and ecological inputs. The use of neural networks in recognising fish stocks is a relatively new development, and it has been restricted to analyses of morphometry or of otolith microchemistry (Murta 2000; Hanson et al. 2004). These were reviewed and summarised by Saila (2005). A more specific introduction to neural network architectures can be found in Dayhoff (1990) and Smith (1993).

These techniques produce continuously distributed values, such as the distance between two points on the body of a fish, or the quantity of a particular element in an otolith. Parasitological studies, on the other hand, are characterised by relatively large number of fish with a low number of parasites, and a small percentage of observations with high numbers. This tends to give problems with classical statistical techniques that require normality assumptions, such as classical discriminant analysis.

Furthermore, some forms of parasite, such as metazoan species, are too numerous to count, therefore a fish is either classed as infected or uninfected. Neural networks are able to cope with both forms of numeric data, as well as with presence–absence data, in any combination.

Note that the question that is addressed by a neural network approach to parasite data is not 'do these fish belong to different stocks', but, 'based on the parasitological data available, is it possible to successfully assign these fish to a stock'? The difference is subtle, but it should be apparent that low levels of successful classification between two sets of observations would not suggest that they belong to two different stocks, whereas high success would support a hypothesis that the samples were drawn from different populations.

25.4 Collection of data

This work is based on samples of *T. trachurus* collected as part of the HOMSIR stock identification project (see Abaunza et al. In press) for a detailed explanation of the theory behind sample collection). Fish were collected with a pelagic trawl by several research vessels at 11 locations in the northeast Atlantic (see Figure 25.1) in 2001 and were immediately frozen and returned to the laboratory for examination. Between 34 and 100 fish were collected from each location (Table 25.1).

To investigate spawning stock identity, three samples each from the Western and Southern stocks, and one each from the North Sea and African stocks, were examined. For estimation of stock mixing, one sample from a non-spawning seasonal fishery from the Norwegian coast and one spawning sample from the boundary between the Western and Southern samples were examined.

Fish were examined externally for parasites, before opening the visceral cavity. All organs were separated, irrigated with physiological saline and examined for the presence of parasites under a stereo-microscope (6–50×). The opercula (gill covers) were removed along with the individual gill arches, irrigated with physiological saline and examined for the presence of monogenean and copepod parasites under a stereo-microscope. Smears of liver and gall bladder were examined for protozoan and myxozoan infections at a magnification of 325× using phase contrast microscopy.

Table 25.1. Location and size of samples collected for stock identification.

Stock	Lat.	Long.	Sample Size
North Sea	54.45N	06.00E	50
Western	52.53N	12.03W	34
	48.45N	09.29W	50
	51.35N	11.06W	50
	44.00N	01.38W	50
Southern	41.00N	08.50W	100
	38.30N	09.20W	52
	37.00N	08.30W	100
African	19.58N	17.28W	50
Mixed Norwegian	57.41N	05.10E	50
Mixed Spanish	43.35N	08.52W	50

25.5 Data exploration

A total of 636 fish from 11 sites were examined. Parasites that infected less than 2% of fish were deemed to represent either rare species or "accidental" infections and were discounted. Eleven species of parasites were found to be commonly present (Table 25.2).

Exploration of any set of data is an essential first step in carrying out an appropriate analysis. One of the problems with this sort of study is that because fish vary in size, age or sex between samples, the examinations cannot be thought of strictly as replications of observations on a homogeneous population. It is therefore important to look for variation caused by these factors and remove it from further analysis.

Plots were made of the abundance of different parasites against fish length, sex and age (Figure 25.4). No relationships were apparent, with the exception of the nematode, *Anisakis* spp. This species encysts in the body cavity of the horse mackerel and therefore is cumulative with age. A $\ln(y + 1)$ transformation was performed on the *Anisakis* abundance values, and a linear regression of the transformed value was carried out against fish length. The residual values of this regression were then taken forwards for use in the later analysis. This is one way of reducing the bias caused by differing lengths between samples.

Table 25.2. Commonly encountered parasite species used for stock identification analysis.

Class	Species	Location	Data Type
Myxosporea			
	Alataspora serenum (Gaevskaya and Kovaleva 1979)	Gall Bladder	Presence/ Absence
Apicomplexa			
	Goussia cruciata (Theolan 1892)	Liver	Presence/ Absence
Nematoda			
	Anisakis spp.	Body Cavity	Abundance
	Hysterothylacium sp. (larval forms)	Body Cavity	Abundance
	Hysterothylacium sp. (adult forms)	Intestine	Abundance
Digenea			
	Tergestia laticollis (Rudolphi 1819)	Intestine	Abundance
	Derogenes varicus (Müller 1784)	Stomach	Abundance
	Ectenurus lepidus (Loos 1909)	Stomach	Abundance
Monogenea			
	Pseudaxine trachuri (Parona and Perugia, 1889)	Gills	Abundance
	Gastrocotyle trachuri (van Beneden and Hesse 1863)	Gills	Abundance
	Heteraxinoides atlanticus (Gaevskaya and Kovaleva 1979)	Gills	Abundance

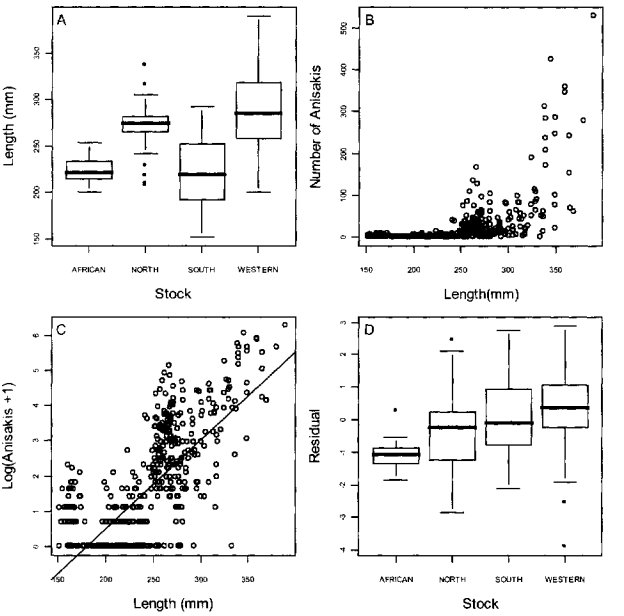

Figure 25.4. A: Fish lengths vary significantly between stock. B: Scatterplot of fish length and *Anisakis* spp. abundance. C: Natural logarithm transformation of *Anisakis* abundance produces a linear relationship. D: The residuals of this relationship are taken forward for use in the analysis after removal of length dependency. The values for the Southern stock are generally lower as the parasite occurred less frequently in this area.

25.6 Neural network results

The first problem in a neural network approach to a classification problem is to select an appropriate structure for the network. This is often done on an *ad hoc* basis. A single hidden-layer, feed-forward neural network was constructed using the *nnet* function (Venables and Ripley 2002) in the R statistical software environment v2.1.1 (R Development Core Team, 2005). This network was provided with half the fish from all samples on a random basis, for use as a training set, then used to reclassify the remaining fish to a stock. In the first instance, the hidden layer contained one object, and to remove chance results caused by selection of fish at random, the process was repeated 100 times with different random training sets. The mean percentage of successful classifications over all simulations is then taken and stored. We then increased the number of units in the hidden layer and repeated the process. Finally, mean successful classification is plotted against number of hidden layer units (Figure 25.5). This allows an educated guess to be made as to the most suitable number of hidden layer objects, which is sufficiently

flexible to reclassify data successfully, but not so many that over-fitting occurs and excessive processing time is required. All other settings remained at their default values.

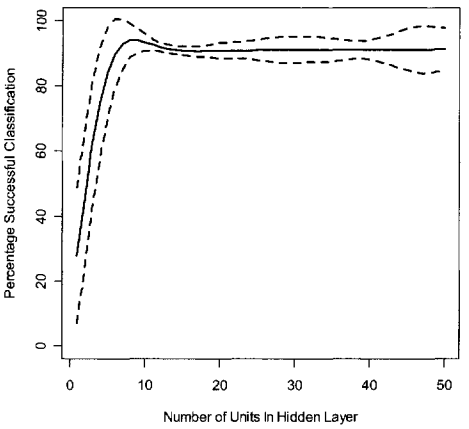

Figure 25.5. Estimation of the optimum number of units in the hidden layer. Mean successful classification (±1 standard deviation) reaches over 90% with eight nodes. Further increases in the number of nodes cause a decrease in success, and an increase in variability, at the cost of increased computing time.

From these data it would appear that the optimum number of hidden layer units is around eight. We decided to use eight hidden layer units for the later neural networks, as there was some decrease in performance of the network at higher values.

To investigate stock identity, the same method of selecting half of the fish in a sample as a training set and reclassifying remaining fish (the "test set") was used. The percentage of fish correctly classified was recorded, along with the numbers from each stock misclassified, and the stock to which they were assigned. This network had 11 inputs, 8 objects in the single hidden layer, and four outputs. To obtain a measure of the error inherent in selecting a training sample at random, and allowing the starting weights of the network to be selected at random, the process was repeated 1000 times. Once outcomes were sorted, median successful classification was represented by the 500^{th} value, and 95% confidence limits by the 50^{th} and 950^{th} values (Figure 25.6).

From these results it is apparent that the neural network is able to correctly classify fish to a stock with a high degree of accuracy — median successful classification for the Southern stock is 95%. These findings support the ICES stock definitions as they are currently applied and suggest that the application of this neural network to mixed stock analysis will give an accurate picture of stock composition. For investigation of the two mixed stock samples, the whole of the spawning dataset was used to train the neural network. This was then used to

reclassify the mixed data in question. The neural network was allowed to choose random starting weights; hence, outputs are still variable, even when considering the same set of data. Consequently, this process was also repeated 1000 times in order to produce an estimate of inherent variability.

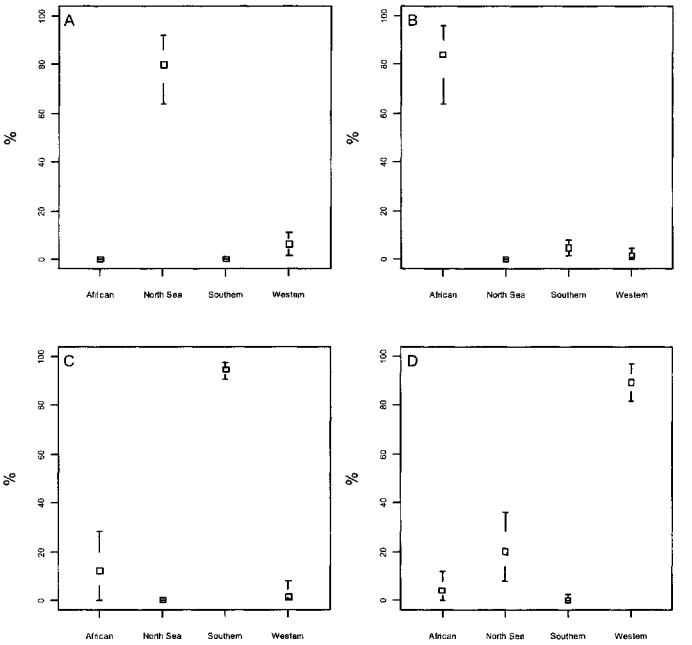

Figure 25.6. Percentage of fish from the test set assigned to each stock by the neural network; A: fish from the North Sea. B: fish from African waters. C: fish from the Southern stock. D: fish collected in the Western stock area.

The Norwegian seasonal fishery shows a more mixed composition than any of the spawning samples, suggesting it may be made up of fish from more than one area. The neural network classifies around 65% of fish as belonging to the Western stock. The remaining 35% are a mixture of Southern and African stocks. Very few fish are assigned to the North Sea stock (Figure 25.7).

The spawning sample from the area of stock uncertainty to the north of Spain is much less conclusive. The neural network assigns around 40% of the sample to the Western stock, 30% to the Southern stock and 20% to the African stock (Figure 25.8).

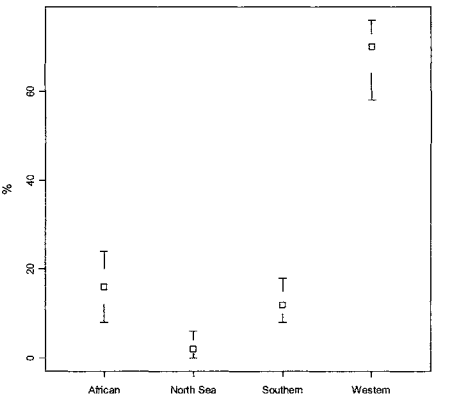

Figure 25.7. Stock membership of horse mackerel from a mixed, non-spawning seasonal fishery that develops in the summer months off the Norwegian coast.

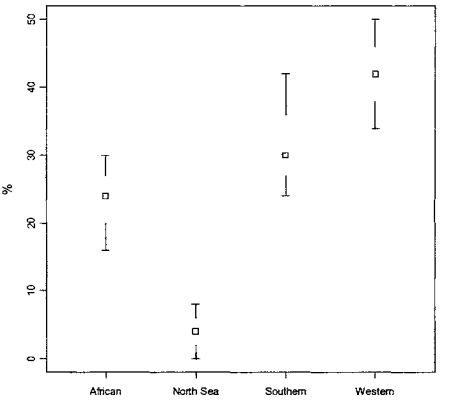

Figure 25.8. Classification of horse mackerel in spawning sample taken from area of stock uncertainty on the northwest coast of Spain.

25.7 Discussion

Stock identification

This study represents the first successful attempt to validate the existence of separate fish stocks by the application of a neural network to parasite abundance data. This technique successfully reclassifies over 90% of fish in the test set, from Western and Southern stocks. Success in reclassifying fish from the North Sea and African stocks is lower, although still over 80%.

The neural network is good at distinguishing between members of the Western and Southern stocks. Median misclassification of fish from the Western stock to the Southern stock is 2%, and from the Southern stock to the Western is 0%. Using parasites as biological tags and multivariate analysis of morphometric distances, it is possible to distinguish between fish from the Western and North Sea stocks, but it has previously been impossible to conclusively distinguish between fish from the Western and Southern areas using any method (ICES 2004).

Although neural networks are robust enough to deal with differences in sample size, it is apparent that the areas with larger sample sizes (western and southern) have higher success rates than the two areas with only 50 fish (North Sea and African stocks). The confidence ranges for these two areas are also much wider than for the larger samples.

It is interesting to note that where misclassification occurs, it tends to be towards stocks that are already believed to mix, rather than to those with which mixing is not regarded as possible. This might suggest that a small number of 'alien' fish from adjacent stocks are present in our 'discrete' spawning samples. These fish would produce such misclassifications, although the difficulties in investigating the processes that go on inside the neural network make this impossible to verify.

It has been suggested that an element of stock mixing can take place between the Western and North Sea stocks while fish over-winter in the English Channel (Macer 1977). This has been supported by recent studies (MacKenzie et al. In press). If stock mixing is occurring in this area, it is likely that this effect is being reflected in the results obtained from the neural network, and that there are a number of fish that belong to the Western stock in the North Sea, and vice-versa. This will slightly confuse the picture that the neural network gives and could explain the tendency of North Sea fish to be misclassified into the Western stock.

A degree of mixing has also been proposed between the Southern and the African stocks (Murta 2000; MacDonald 2005). This is supported by our findings. Although successful classification of fish from the Southern stock is over 95%, the neural network classifies around 20% of fish from the African sample as belonging to the Southern stock. Very few fish from the southern samples are classified as belonging to the African stock. This could suggest that mixing between these areas is a one-way process.

Norwegian non-spawning sample

Having established the utility of a neural network approach to assign fish to spawning stocks based on host-parasite data, it is a straightforward matter to apply this to a non-spawning stock to investigate its composition. These findings suggest that fish in the Norwegian sample do not come from a single stock, but rather are drawn mainly from the Western stock, with a sizeable proportion from much more southerly stocks. This finding is in line with work by Abaunza (In press) who measured growth rates of horse mackerel and found variability from stock to stock, with fishes from warmer waters growing more quickly than their more northerly conspecifics. When examining fish from the Norwegian area, growth

rates were noticeably higher than that in neighbouring fisheries, to the west of Ireland and the southern North Sea. Abaunza proposed the existence of a highly migratory 'infra-stock' that spawned and over-wintered in the Southern stock area, then migrated northwards to feed in Norwegian waters. Evidence supporting this hypothesis came from the discovery of characteristically "southern" parasites in fish from the Norwegian area (MacKenzie et al. In press).

The neural network classified few fish from this sample as belonging to the North Sea stock. This is interesting when considering how close it is to this stock. It suggests that very little mixing of North Sea and other fish occurs in this area. This finding is important for fishery managers to consider when allocating catches of fish from the Norwegian fishery to particular spawning populations.

La Coruña spawning sample

The neural network was not able to classify fish from the north west coast of Spain to a particular stock with any great certainty. No one stock appears to dominate this area, and the 95% confidence limits are relatively small. The boundary between the Western and the Southern stocks was recently moved from the north to the west coast of the Iberian peninsula (ICES 2004). These findings suggest that this change was appropriate, in that the sample is classed as more "Western" than 'Southern', but also suggest that a high degree of mixing takes place in this area, and that more intensive sampling in this area would be a worthwhile contribution to stock identification of the horse mackerel.

Conclusions

It is apparent that a neural network approach to classifying individuals into presupposed groups is a powerful tool for problems such as this.

The ease with which it is possible to use neural networks, their lack of restrictive assumptions and their ability to cope with combinations of different types of data make them extremely useful for dealing with problems of classification in ecology.

Acknowledgements

We are grateful to all our partners in the HOMSIR stock identification project for their helpful advice and provision of samples. The work that this chapter was based on was funded by the European Commission, under the 5th Framework, contract no. QLRT-PL1999-01438. Visit www.homsir.com for more details on the project outcomes. We would like to thank Anatoly Saveliev for valuable comments on the statistical aspects of this chapter.