24 Classification trees and radar detection of birds for North Sea wind farms

Meesters, H.W.G., Krijgsveld, K.L., Zuur, A.F., Ieno, E.N. and Smith, G.M.

24.1 Introduction

In Chapter 9, we introduced univariate regression trees and briefly discussed classification trees. In this chapter we expand on Chapter 9 and provide a detailed explanation of classification trees applied to using radar records to identify bird movements important in choosing sites for offshore wind farms. Tree model software tends to produce large amounts of numerical output, which can be difficult to interpret, and this chapter provides a detailed discussion on interpreting this output.

To increase the supply of renewable energy in the Netherlands, the Dutch government is supporting the construction of a 36-turbine Near Shore Wind farm (NSW) located 10–15 km off the coast of Egmond in the Netherlands (Figure 24.1). This project serves as a pilot study to build knowledge and experience of the construction and exploitation of large-scale offshore wind farms. An extensive monitoring and evaluation programme has been designed to gather information on economic, technical, ecological and social effects of the NSW. This evaluation programme will give information for future offshore wind farm projects as well as an assessment of the current NSW.

Derived from land-based studies, the ecological monitoring programme requires an analysis of three types of possible effects of wind farms on birds: the collisions of birds with turbines, the disturbance of flight paths and possible barrier effects and the disturbance of resting and feeding birds. The project discussed here focused on flight paths, mass movements and flight altitudes of flying birds. It was carried out by Bureau Waardenburg and Alterra Texel and was commissioned by the Dutch National Institute for Coastal and Marine Management. The full results of the study are published in Krijgsveld et al. (2005).

To assess the risks from collision and disturbance of flight paths on birds, it is necessary to identify and quantify the flight patterns of birds in the area, prior to building of the wind farm. Flight patterns were quantified using a combination of automated and field observation techniques. All birds flying through the study area were recorded by means of an automated system using two radars that processed and stored signals in two databases.

One radar rotated horizontally and recorded the direction and speed of all birds flying through the study area. Another radar rotated vertically and recorded the movement and altitude of birds flying through an imaginary line suspended vertically above the radar. This automated system was operated continuously and collected flight data every day of the year, both day and night. In addition, field observations were used to obtain detailed information on bird species composition and behaviour to validate the automated measurements. These field observations were made from an observation platform at sea, Meetpost Noordwijk (Figure 24.1), and comprised observations of mass movements, flight paths and flight altitudes of birds during the day and to a lesser extent during the night. The two radars operated from the same location and were equipped with software designed to distinguish bird echoes from other types of echoes, such as ships and clutter (echoes resulting from radar energy reflected by waves, clouds or other atmospheric conditions), but this was ineffective as a lot of the clutter was still recorded.

It is obviously important to be able to distinguish bird echoes from non-bird echoes if we hope to get an accurate picture of bird movements at sea. We therefore investigated how the characteristics recorded for each echo could be used to identify different classes of birds, or at least help to distinguish echoes from birds from echoes of clutter and ships. In this chapter we show how classification trees were used to classify the echoes from the horizontal radar.

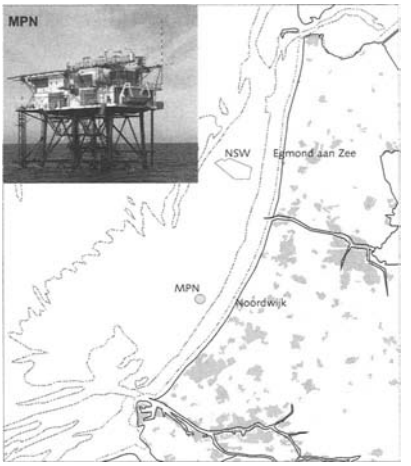

Figure 24.1. Location of planned Near Shore Wind park (NSW) and the observation platform, Meetpost Noordwijk (MPN), off the west coast of the Netherlands. Top left is a picture of the observation platform.

24.2 From radars to data

Observation platform 'Meetpost Noordwijk' is situated at approximately the same distance from the coast as the NSW area (Figure 24.1). It has three decks,

the highest of which is 19 m above sea level, giving a good height to observe birds. Two types of radars were used. One, for the observation of flight paths, was a 30-kW Furuno S-band horizontal surveillance radar. This is a standard radar, as used on ships, and scans the area in the horizontal plane around the radar. With this radar, the flight paths of birds flying through the radar beam were tracked and flight speeds (ground speed) and directions were recorded. The second type of radar was a 25-kW Furuno X-band surveillance radar tilted vertically, which was used for the observation of mass movements of birds and flight altitudes. Both radar systems with associated software were installed on the observation platform to continuously measure all bird movements in an area up to 11 km from the platform in every horizontal direction and up to 3 km above the platform. These radars scanned the area above the sea continuously throughout the year, both day and night, and automatically recorded the exact location, direction, speed, and altitude of all flying objects within the scanned area. These data provided the principle dataset on flight patterns, and far more extensive than the field observations because it is continuous, it can record at night and can still record bird movements in poor weather conditions.

In this chapter we only discuss the analyses of data from the horizontal radar. Echoes originate not only from birds, but also from ships, planes, helicopters, or can just be reflections (clutter) from waves, termed wave/sea clutter, or from atmospheric phenomena, called air clutter. The radar equipment records and stores all echoes that it detects unless filtered out by the software as belonging to objects other than birds. Although this should only leave birds in the database, a large percentage of stored echoes still belong to other objects, mainly waves. To separate bird echoes from echoes of other objects (ships, clutter etc), the echoes need to have key identifying characteristics that can be assigned to a certain group, *a posteriori*. To help identify these key characteristics, a fairly large number of characteristics were recorded for each echo (Table 24.1).

Determining the characteristics of radar echoes from different objects requires a training set, a dataset of stored echoes from known objects and these were assembled during fieldwork at Meetpost Noordwijk. During training sessions, field observers and radar operators were in radio contact, so that radar detected objects could be located by the observers and the relevant object visually identified. The resulting set of verified radar echoes were analysed by classification trees to obtain relationships by which *a posteriori* radar images could be identified as belonging to certain groups. The software was set up to follow the flight path of each bird (object) or group of birds and record this as one track. If a bird was lost during several rotations of the radar, for example, due to a sudden movement of the bird, or because it landed on the water, the bird may have been recorded in a second track as well. We assumed that this would generally be the exception in our data and that records (i.e., tracks) were uncorrelated and were from different birds.

The original data consisted of echoes with their associated characteristics as stored by the radar software, plus information from the observer who verified the type of object associated with that the echo. This information was divided into different categories:

- General information (e.g., date, time).

- Information recorded by the observer at the same moment as the echo was made (e.g., species, flock size, flight altitude).
- Echo appearance information (e.g., echo dimensions, reflectivity).
- Echo position (e.g., x and y coordinates in the radar plane, distance from radar).
- Echo movement (e.g., speed).

Some of these variables represent the same ecological information and had high (>0.8) correlations. Using common sense and statistical tools like correlations and principal component analysis, we condensed the number of original variables into a subset of important characteristics (Table 24.1).

Table 24.1. Variables and abbreviations used in the analyses.

Variable	Description
EPT	Echoes per track.
TKQ	Track quality defined as STT/EPT. STT is the sum of all the TKT within 1 Track..
TKT	Track type, measure for the consistency with which the track was seen by the radar.
AVV	Average velocity of object based on all echoes of one track.
VEL	Velocity of object.
MXA	Maximum echo area (in pixels) of all echoes belonging to one track.
AREA	Area of the target in pixels.
MAXREF	Maximum reflectivity of all echoes in a track.
TRKDIS	Distance covered by the whole track.
MAXSEG	Longest length across the target.
ORIENT	The angle of the longest axis of a target with respect to the horizontal axis. This value is between 0 and 180 degrees.
ELLRATIO	Ratio of Ellipse Major to Ellipse Minor. Ellipse Major/Minor is the length of the major/minor axis of an ellipse that has the same area and perimeter as the target.
ELONG	A measure of the elongation of a target, the higher the value the more elongated the target.
COMPACT	Compactness, defined as the ratio of the target's area to the area of the smallest rectangle.
CHY	The mean length, in pixels, of the vertical segments of a target.
MAXREF	Maximum reflectivity over the entire target area.
MINREF	Minimum reflectivity over the entire target area.
SDREF	Standard deviation in reflectivity over the entire target area.

24.3 Classification trees

Univariate regression and classification trees (Chapter 9, Breiman et al. 1983; Therneau 1983; De'ath and Fabricus 2000) can be used to model the relationship

between one response variable and multiple explanatory variables. The tree is constructed by repeatedly splitting the data using a rule based on a single explanatory variable. At each split the data are partitioned into two mutually exclusive groups, each of which is as homogeneous as possible. Splitting is continued until an overly large tree is grown, which is then pruned back to the desired size. This is equivalent to the model selection procedure in linear regression (Chapter 5). Tree models can deal better with non-linearity and interaction between explanatory variables than regression, GLM, GAM or discriminant analysis. Classification trees are used for the analysis of a nominal response variable with two or more classes, and regression trees for a non-nominal/numeric response variable. With classification trees, a transformation of the explanatory variables does not affect the results.

The classification tree tries to assign each observation to one of the predefined groups based on a specific value from one of the variables, thereby maximizing the variation between the groups while minimizing the variation within each group.

The main problem with tree models is determining the optimal tree size: A full tree with lots of splitting rules is likely to overfit the data, but a tree of size of only two or three might give a poor fit. The process of determining the best tree size is called 'pruning the tree'. An AIC type criterium is used to determine how good or bad is the tree. Recall from Chapter 5 that the AIC was defined as

$$\text{AIC} = \text{measure of fit} + 2 \times \text{number of parameters}$$

As measure of fit we used the total sum of residual squares in linear regression and the deviance in GLM and GAM. For a tree model, we have

$$\text{RSS}_{\text{cp}} = \text{RSS} + \text{cp} \times \text{size of tree} \tag{24.1}$$

Where RSS stands for residual sum of squares and cp is a constant. The size of the tree is defined as the number of splits plus one. The RSS component in equation (24.1) is equivalent to the measure of fit in the AIC, size of the tree equivalent to the number of parameters and cp to the × 2 multiplier. In linear regression, we just change the number of parameters and select the model with the lowest AIC. Here, things are slightly more difficult as we do not know the value of cp. If we knew the cp value, then we could do the same as with the AIC and just calculate the RSS_{cp} for different sizes of the tree and choose the tree with the lowest RSS_{cp} value. However, as we do not know the cp value, we need to estimate it, together with the optimal tree size, and use a cross-validation process, which is discussed later. First, we need to discuss how to calculate the residual sum of squares.

Consequently, the next two paragraphs are slightly more technical and the rest of this section may be skipped by readers not interested in this.

Estimating RSS in equation (24.1)

It is easiest to look at 0–1 data first. A leaf represents a group of observations that are deemed similar by the tree and are plotted at the end of a branch. The RSS component, also called the deviance D, at a particular leaf j is defined as

$$D_j = -2[n_{1j} \log \mu_j + n_{0j} \log(1 - \mu_j)]$$

where n_{1j} is the number of observations in leaf j for which $y = 1$ and n_{0j} is the number of observations for which $y = 0$. The fitted value at leaf j, μ_j, is the proportion $n_{1j}/(n_{1j} + n_{0j})$. The overall deviance of a tree is the sum of the deviances over all leaves. If the response variable has more than two classes (e.g., five), the deviance at leaf j is defined as

$$D_j = -2\sum_{i=1}^{5} n_{ij} \log \mu_{ij}$$

$$D_j = -2[n_{1j} \log \mu_{1j} + n_{2j} \log \mu_2 + n_{3j} \log \mu_{3j} + n_{4j} \log \mu_{4j} + n_{5j} \log \mu_{5j}]$$

where n_{1j} is the number of observations at leaf j for which $y = 1$, and $\mu_{1j} = n_{1j}/(n_{1j} + n_{2j} + n_{3j} + n_{4j} + n_{5j})$.

Estimating cp in equation (24.1)

The parameter cp in equation (24.1) is a constant. For a given value, the optimal tree size can be determined in a similar way to choosing the optimal number of regression parameters in a regression model. Setting $cp = 0$ in equation (24.1) results in a very large tree as there is no penalty for its size. The other extreme is $cp = 1$, which will result in a tree with no leaves. To choose the optimal cp value, cross-validation can be applied. With this approach, if the data are split up into, say, 10 parts, a tree is fitted using data of 10 parts, and the tenth part is used for prediction. The underlying principle of this approach is simple; leave out a certain percentage of the data and calculate the tree. Once the tree is available, its structure is used to predict in the group that the omitted data belongs to. As we know which groups the omitted data belong, the actual and predicted values can be compared, and a measure of the error calculated: the prediction error. This process is applied for each of the $k = 10$ cross-validations, giving 10 replicate values for the prediction error. Using those 10 error values, we can calculate an average and standard deviation (an illustration of this process is given later). The entire process is then repeated for different cp values in a 'back-ward selection type' approach. Examples are provided in the next two sections.

24.4 A tree for the birds

The verified dataset consists of 659 cases divided over 16 groups (Table 24.2). In this section we try to classify 9 different groups with at least 10 observations resulting in 629 observations (this excludes cormorants, gannets, land birds, skuas, unidentified birds, ducks, and waders). The nine groups in the analysis were auks, air clutter, water clutter, divers, geese and swans, gulls, sea ducks, ships and terns.

The graphical output of the classification tree for these data is given in Figure 24.2 and Figure 24.3.

We discuss the cross-validation graph (Figure 24.2) first. This graph was obtained with a default value of $cp = 0.001$. The average and the standard deviation of the 10 cross-validations are plotted versus cp and the tree size. The complexity parameter is labelled along the lower x-axis, and the tree size is labelled along the upper x-axis. The y-axis gives the relative error in the predictions, obtained by cross-validation, and the vertical lines represent the variation within the cross-validations (standard deviation). This graph is used to select the most optimal cp value. A good choice of cp is the leftmost value for which the mean (dot) of the cross-validations lies below the horizontal line. This rule is called the one standard deviation rule (1-SE). The dotted line is obtained by the mean value of the errors (x-error) of the cross-validations plus the standard deviation (X-std) of the cross-validations upon convergence. Figure 24.2 indicates that a tree with one split (size of tree = 2) would be the best size of tree. Little would be gained from a tree with more splits.

The final tree, calculated with a *cp*-value of 0.0323 (as suggested by the lowest error in Figure 24.2), is presented in Figure 24.3. The tree is arranged so that the branches with the largest class go to the right. Branch length is proportional to the improvement in the fit. Below each branch the predicted class (or group) and the number of observations in each class are given. The results show that both types of clutter are completely grouped in the left branch predicting that records will be from wave or air clutter (class = 3). Note that this branch includes four ships. The other branch includes all the birds, and as class seven (gulls) had the largest number of cases, all echoes from gulls are included in this leaf. Apart from the wrongly classified non-gulls, 30 ships are also classified as gulls. The variable that can best be used for the first (and only) split is EPT, the number of echoes per track. As EPT are integers, this means that everything with a single EPT is classified as clutter and the rest is classified with the gulls.

Table 24.2. Number of echoes per group. Groups in italics were not used in the tree analysis, which aimed at separating the 9 groups with more than 10 observations per group.

Group	Number of Obs.	Group	Number of Obs.
Auks	11	*Land birds*	8
Clutter air	37	Sea ducks	18
Clutter water	52	Ship	34
Cormorants	2	*Skuas*	1
Divers	22	Terns	16
Gannets	5	*Unidentified Bird*	7
Geese and swans	27	*Unidentified duck*	5
Gulls	412	*Waders*	2

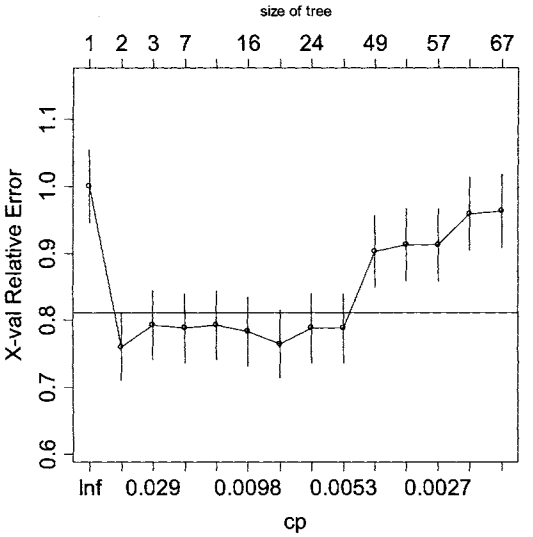

Figure 24.2. Cross-validation results for radar data separated into nine groups. Optimal tree size is 2.

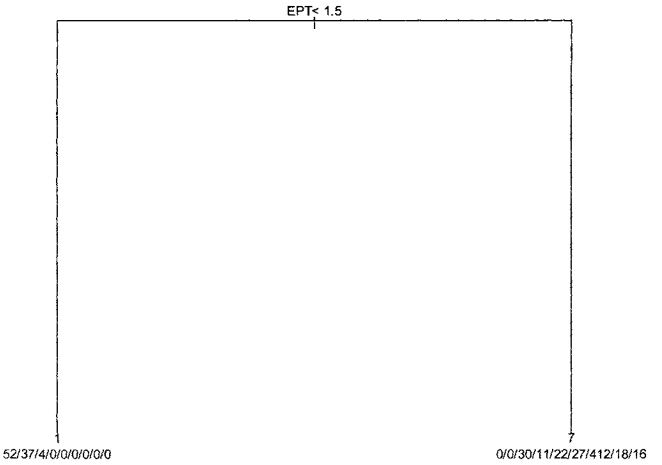

Figure 24.3. Tree for nine groups of radar data using a cp setting of 0.0323. Principal division is based on Echoes per Track (EPT). Groups are from 1 to 9: water clutter, air clutter, ships, auks, divers, geese and swans, gulls, sea ducks, and terns. Branch length is proportional to the improvement in the fit. Below each branch the predicted class is given first, followed by the number of observations in each class.

The numerical output of the tree

Readers not interested in the explanation of the numerical output might want to skip this paragraph in their first reading of this chapter. A lot of numerical output is produced by tree software, and some of the output is presented in Table 24.3. This printout shows that the cross-validation mean value at the largest tree is 1.00 (this is a percentage of the root node error), and the standard deviation is 0.055, which added together gives 1.055. Normally, this value would be taken for the 1-SE rule (representing the horizontal line in Figure 24.2); however, in this case, the tree is not decreasing regularly and the value used for the 1-SE rule is the lowest error value and its standard deviation (0.76 + 0.051 = 0.81) is already reached after the first split. The smallest tree that has a smaller mean cross-validation error (0.76) has one split and is therefore of size two with a *cp*-value of 0.0323 (Figure 24.2). The error columns have been scaled so that the first node has an error of one (multiply columns 3–5 by 217 to get a result in terms of absolute error).

The root node error for a classification tree is the classification error before any splits have been made. Because most samples in the dataset were from group seven (412 gull records), the algorithm classified the entire dataset as group seven. Therefore, samples of all other groups, 217 in total, are wrongly classified, and the root node error is 217 out of 629 (= total number of samples), which is 0.34. Using one split, which corresponds to a tree size of two, results in an error of 76% of the root node error. A tree of size five (Nsplit = 6) has an error of 63% of the root error (Table 24.3).

Table 24.3. Results of cross validation. cp, complexity parameter; nsplit, number of splits; rel error, relative error in the predictions; xerror, the mean value of the errors of the cross-validations; xstd, the standard deviation of the cross-validations. Root node error: 217/629 = 0.34.

	Cp	Nsplit	Rel-error	x-error	x-std
1	0.2396	0	1	1	0.055
2	0.0323	1	0.76	0.76	0.051
3	0.0253	2	0.73	0.78	0.051
4	0.0138	6	0.63	0.78	0.051
5	0.0104	9	0.59	0.78	0.051
6	0.0092	15	0.51	0.81	0.052
7	0.0069	21	0.46	0.82	0.052
8	0.0061	23	0.44	0.83	0.052
9	0.0046	26	0.42	0.83	0.052
10	0.0037	48	0.32	0.89	0.053
11	0.0031	53	0.3	0.94	0.054
12	0.0023	56	0.29	0.94	0.054
13	0.0018	58	0.29	0.97	0.055
14	0.001	66	0.27	1.00	0.055

The numerical output for the final tree constructed with the *cp*-value of 0.0323 (Figure 24.2) was:

```
Node number 1: 629 observations, complexity param = 0.24
  predicted class = 7  expected loss = 0.34
  class counts:   52   37   34   11   22   27  412   18   16
  probabilities: 0.083 0.059 0.054 0.017 0.035 0.043 0.655 0.029 0.025
  left son = 2 (93 obs) right son = 3 (536 obs)
  Primary splits:
     EPT   < 1.5 to the left,  improve=85, (0 missing)
     TKT   < 2.5 to the right, improve=49, (0 missing)
     MAXREF < 940 to the left,  improve=44, (1 missing)
     MXA    < 8.5 to the left,  improve=43, (0 missing)
     AVV    < 57  to the right, improve=29, (0 missing)
  Surrogate splits:
     MXA    < 8.5 to the left,  agree=0.933, adj=0.548, (0 split)
     MAXREF < 940 to the left,  agree=0.930, adj=0.527, (0 split)
     TKT    < 2.5 to the right, agree=0.908, adj=0.376, (0 split)
     TKQ    < 3.9 to the right, agree=0.879, adj=0.183, (0 split)
     AVV    < 66  to the right, agree=0.876, adj=0.161, (0 split)

Node number 2: 93 observations
  predicted class=1  expected loss=0.44
  class counts:   52   37    4    0    0    0    0    0    0
  probabilities: 0.559 0.398 0.043 0.000 0.000 0.000 0.000 0.000 0.000

Node number 3: 536 observations
  predicted class=7  expected loss=0.23
  class counts:    0    0   30   11   22   27  412   18   16

  probabilities: 0.000 0.000 0.056 0.021 0.041 0.050 0.769 0.034 0.030
n= 629
node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 629 217 7 (0.083 0.059 0.054 0.017 0.035 0.043 0.66 0.029 0.025)
   2) EPT< 1.5 93  41 1 (0.56 0.4 0.043 0 0 0 0 0 0) *
   3) EPT>=1.5 536 124 7 (0 0 0.056 0.021 0.041 0.05 0.77 0.034 0.03) *
```

The first node is the root of the tree, representing the undivided data, and has 629 observations. The complexity parameter indicates that any *cp*-value between 0.0323 (as suggested in Figure 24.2) and 0.24 would have given the same result. The predicted class for the first split is based entirely on the number of observations in each class. As class seven (gulls) has the highest number of observations the predicted outcome for the first split is class seven. As there are 217 observations in other classes the expected loss (samples incorrectly classified) is 217/629 = 0.34. After this the number of observations in each class together with the relative probabilities for each class are given, followed by the number of observations in the left and right part of the split. Then, the variables available for each split are

given. These are ordered by the degree of improvement, with the variable that results in the highest classification score presented first. The actual values of the improvement are not so important, but their relative size gives an indication of the comparative importance of the variables. Clearly, using variables other than EPT results in a serious loss of fit. The primary variables are followed by surrogate variables. These are variables that can be used instead of the first primary variable for cases where a value for the primary variable is missing. The percentage of agreement with the classification of the primary split (for both directions of the split) is given under "agree" (0.933 for MXA). Next, 'adj' gives the adjusted concordance for surrogate splits with the primary split, meaning how much is gained beyond 'go with the majority; rule (calculated as adj = (agree − 536/629)/(1 − 536/629)).

At the end of the detailed numerical output, a summary of the tree analysis is given based on the best variable for each split. The root is always node number one, and the following nodes are defined as twice the 'previous-node-number' for the left split, and twice the 'previous-node-number'+1 for the right split. The splitting rule is given ('split'), the number of cases (n), the number that does not follow the rule and thus are incorrectly classified (loss), the predicted class (yval), and the numbers in each class as a fraction of the total (yprob).

Summary

Classification tree analysis was used to find out how to distinguish among nine different groups of echoes originating from air clutter, wave clutter, ships, auks, divers, geese and swans, gulls, sea ducks, and terns. The tree cannot discriminate among the different groups of birds and wrongly classifies 88% of the ships as birds. However, it classifies 100% of the clutter and 100% of the birds correctly. Because the groups are rather diverse and the analysis indicated that we could not reliably classify all nine groups, we re-applied the tree analysis on the same data but this time grouping them into four group: Birds, ships, air clutter and wave clutter. This is discussed in the next section.

24.5 A tree for birds, clutter and more clutter

For this analysis we grouped all birds into a single group. This gave four groups for the tree to classify: air clutter in air, wave clutter ships and birds. The resulting classification tree is presented in Figure 24.4. The cross-validation plot (Figure 24.5) indicates that a tree with two branches is the best, but this still classifies ships and birds as belonging to the same group. To see whether there was a variable that could separate ships from birds, we used a *cp* value for the tree that was smaller than suggested by the *cp*-plot in Figure 24.5. The resulting tree (Figure 24.4) indicates that by using an extra variable, namely AREA, 41% of the ships can be correctly identified at the cost of misclassifying 2 of the 506 birds. In total, only 3.5% of the data were wrongly classified (taking the two types of clutter to-

gether). The tree also indicates that most birds can be discriminated from ships by using AREA (area of the target in pixels). Evidently, ships have quite a different size to the average bird, and this is reflected in the number of pixels on the radar screen. Birds that overlap in area with ships may be (partly) separated by TKQ, track quality, which is generally smaller for birds.

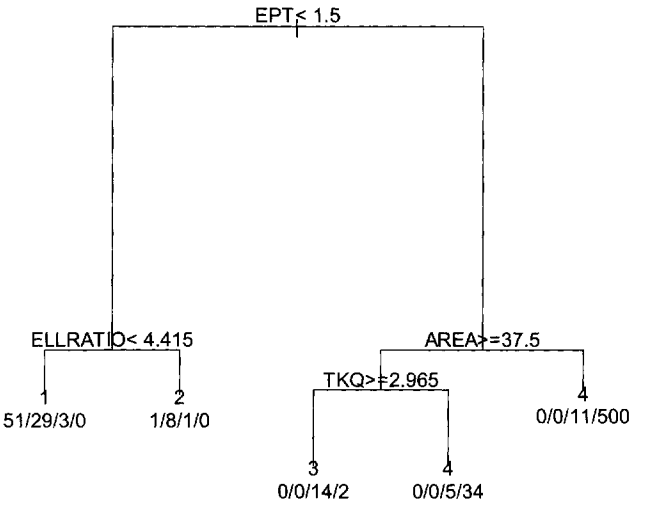

Figure 24.4. Tree for horizontal radar data using four groups and a cp setting of 0.035. Groups are from 1 to 4: air clutter, water clutter, ships, and birds. Branch length is proportional to the improvement in the fit. Below each branch first the predicted class is given, followed by the number of observations in each class.

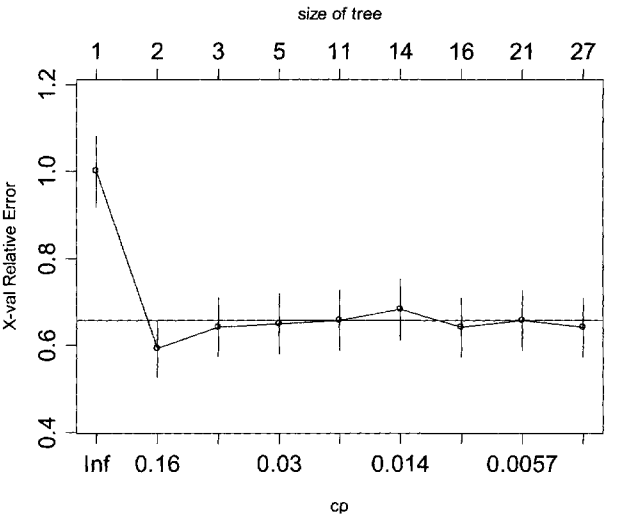

Figure 24.5. Cross-validation results for radar data separated into four groups. Optimal tree size is 2.

A summary of the tree analysis of horizontal radar data using a *cp* value of 0.035 is given below. A '*' denotes a terminal node. The classes in yprob are air clutter, wave clutter, ships and birds.

node)	split	n	loss	yval	(yprob)
1)	root	659	123	4	(0.078, 0.056, 0.051, 0.813)
2)	EPT< 1.5	93	41	1	(0.559, 0.397, 0.043, 0.000)
4)	ELLRATIO< 4.42	83	32	1	(0.614, 0.349, 0.036, 0.000) *
5)	ELLRATIO>=4.42	10	2	2	(0.100, 0.800, 0.100, 0.000)*
3)	EPT>=1.5	566	30	4	(0.000, 0.000, 0.053, 0.947)
6)	AREA>=37.5	55	19	4	(0.000, 0.000, 0.345, 0.654)
12)	TKQ>=2.965	16	2	3	(0.000, 0.000, 0.875, 0.125)*
13)	TKQ< 2.965	39	5	4	(0.000, 0.000, 0.128, 0.871)*
7)	AREA< 37.5	511	11	4	(0.000, 0.000, 0.021, 0.978)*

24.6 Discussion and conclusions

Trees can be used for both description and prediction. Statistical analysis of data is often separated into two phases: exploration and modelling, with the former preceding the latter. Trees can be used for both as shown in this chapter. As an exploratory tool, trees create structure in the data and lead to model building. As a modelling tool, trees may generate models that represent the systematic structure of the data as simply as possible and they can also be used as a prediction tool by accurately predicting unobserved data.

We have shown how classification trees can be used to split echo images from radar at a platform in the North Sea into different classes. Aiming high we first tried to separate the data into nine different classes, which included clutter from waves, clutter from air disturbances, ships and six classes of birds. Starting with 17 different radar variables, the first analysis generated an optimal solution with only one variable: the number of echoes per track. For the sampled data the tree accurately distinguished the two clutter groups from the rest, but was unable to separate the different bird classes from ships nor to distinguish between the two forms of clutter. This led us to try a different approach of lumping all the bird groups into one class called 'birds'. This new analysis again suggested a tree with only one split, but increasing the tree size to one with four splits allowed us to separate a large proportion of the ships from the birds. This new tree correctly classified 96.5% of the data (with wave clutter and air clutter put into one class because our goal was mainly to get rid of the clutter regardless of its origin). Birds (and ships) were mostly followed for several radar rotations and generally had more echoes per track than clutter. This led to the number of echoes per track (EPT) being the main variable in the trees (also indicated by the greater length of the branches following the EPT split, (Figure 24.3 and Figure 24.4). However, there appeared to be a substantial amount of overlap between the different variables, which made it impossible to separate all ships from birds. However, the tree (Figure 24.4) indicates that area can sometimes be used to differentiate ships from birds, but with 36 birds with areas comparable with ships. It is difficult to reconcile how birds can be as big as ships, but looking closely at the data we found that of these 36, 17 consisted of more than one individual (2 to 750 birds), indicating a possible effect of flock size. In addition, AREA is measured as pixels on the radar screen, and depending on the detection distance and detection limit of the radar, an echo of a bird or a ship may be indicated with an identical number of pixels. Using track quality brought about some further improvements, but there remained a small overlap between birds and ships. Also in the branch with smaller areas, a number of ships can be found. This is also clear from dotplots (not shown here).

In conclusion tree analysis proved itself a useful tool to separate clutter from the rest of the data, but it did not succeed fully in separating birds from ships.

Acknowledgement

We would like to thank S. Dirksen, R. Lensink, M.J.M. Poot, and P. Wiersma of Bureau Waardenburg, and H. Schekkerman of Alterra for their help at various stages of the research. The radars and accompanied software were supplied by DeTect Inc. (Florida, USA). We would also like to thank Neil Cambell for comments on an earlier draft.