21 Analysing presence and absence data for flatfish distribution in the Tagus estuary, Portugal

Cabral, H., Zuur, A.F., Ieno, E.N. and Smith, G.M.

21.1 Introduction

Understanding the spatial and temporal distribution and abundance patterns of species and their relationships with environmental variables is a key issue in ecology. All too often, despite best efforts, the quality of the collected data forces us to reduce it to presence–absence data, which presents the ecologist with several specific statistical problems. In this chapter, using fish data as an example, we look at several approaches for analysing presence–absence data.

Several statistical techniques have been used to relate fish abundance with abiotic and biotic conditions. These include regression methods, general linear models, ordination methods, discriminant analysis, and several others (Jager et al. 1993; Eastwood et al. 2003; Thiel et al. 2003; Amaral and Cabral 2004; França et al. 2004). However, methods that evaluate the performance of different techniques applied to the same datasets are scarce. And the choice of analytical approach can strongly influence the statistical conclusions and, by implication, the ecological conclusions. This case study compares the adequacy and the performance of several statistical tools when used to analyse fish abundance data and its relationships with environmental factors.

The data used in this example are the abundance of sole, *Solea solea* in the Tagus estuary (Portugal), with the principal ecological question of identifying which environmental factors influence the choice of nursery grounds by this species, in this estuary. The sole is a marine fish that occurs and spawns throughout the continental shelf. Larvae and juveniles tend to migrate to coastal nursery areas using both passive and active transport processes (e.g., Rijnsdorp et al. 1985). The juveniles concentrate in estuaries and bays for a period of about two years (Koutsikopoulos et al. 1989).

The Tagus estuary (Figure 21.1) has long been recognized as an important nursery area for *S. solea*. This estuarine system, located in the centre of the west coast of Portugal, has an area of 325 km^2 and is a partially mixed estuary with a tidal range of about 4 m. The Tagus estuary has long been subjected to industrial development, urbanization and port and fishing activities. Over two million people live around the estuary, mainly in the lower part, where, important industrial complexes such as chemical, petrochemical, food and smelting are found. The upper part (more than 50% of the shoreline) is bordered by land used intensively for agriculture (Fernandes et al. 1995). The *S. solea* abundance data used in this case study were collected during monthly sampling surveys conducted in four areas during 1995 and 1996. Fish were captured using a 4 m beam trawl with 10 mm mesh and one tickler chain. Sampling areas were selected for their importance to juvenile sole, based on results from earlier studies (Costa and Bruxelas 1989). Several environmental variables were also measured: depth, salinity, water transparency, temperature and sediment composition (percentages of mud, medium and fine sand, large sand and gravel).

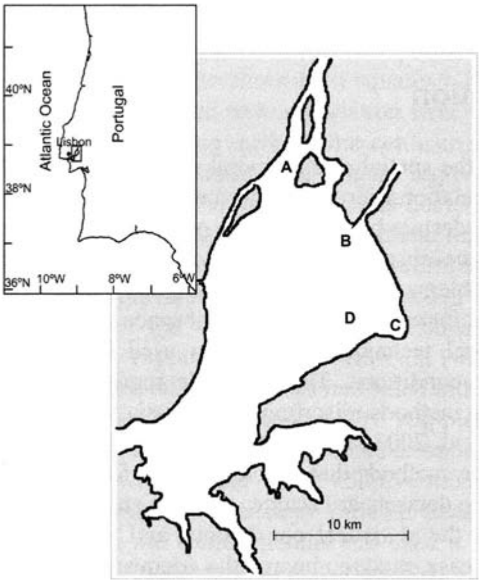

Figure 21.1. The sampling area in the Tagus estuary, Portugal. A, B, C and D indicate the location of the stations.

21.2 Data and materials

The *S. solea* data used in this chapter consist of density values (numbers of individuals per 1000 m^2), recorded from 65 samples in 1995 and 1996. The samples were collected from four areas (Figure 21.1). Sixty percent of the *S. solea* samples were zero, and a few observations had high abundance. This extreme distribution of the data makes it difficult for analysis, as it does not match many models available from standard statistical tools — a common problem with abundance data. It is too extreme for transformation to be successful, and although several methods such as generalised linear models (Chapter 6) have the capacity to deal with overdispersed data, there is a limit to the amount of overdispersion that can be compensated for. In this instance, we decided there was little choice but to reduce it to presence and absence data. Accept the loss of the quantitative information for the sample points where the fish were present. This makes the presence or absence of fish the nominal response variable we are trying to explain using a range of available explanatory variables.

Table 21.1 shows the available explanatory variables. The variables season, month and station are nominal. From the statistical point of view, we have only one response variable, the presence or absence of *S. solea*, and multiple explanatory variables. The question is whether there is a relationship between the two. Because of the presence–absence nature of the response variable, the most appropriate techniques for this analysis are generalised linear models (GLM), generalised additive models (GAM) using a binomial distribution, and classification trees. The number of explanatory variables is more than 10 (8 continuous and 3 nominal). This can be considered as a large number of explanatory variables and might make it difficult to find the optimal model, especially with explanatory variables like mud and the fine sand percentage of the sediment, which have a degree of collinearity. To simplify the problem, our strategy was to:

1. Identify and remove some of the highly correlated (continuous) explanatory variables. Pairplots are a useful tool for this task.
2. Visualise the relationship between the nominal explanatory variables and *S. solea*. Design and interaction plots will be used for this.
3. Visualise the relationship between *S. solea* and the explanatory variables by using coplots.
4. Use classification trees, GAM and GLM to model the relationship between *S. solea* and the selected explanatory variables, and compare the results among all three techniques.

The selection of the explanatory variables in the first step should be based on ecological and statistical considerations.

Table 21.1. Available explanatory variables.

Variable	Nominal	Remark
season	Yes	1 = spring, 2 = summer
month	Yes	
station	Yes	sampling station
depth	No	Depth (m)
temp	No	temperature (°C)
sal	No	salinity (ppt)
transp	No	water transparency (cm)
gravel	No	% gravel in the sediment
large sand	No	% large sand in the sediment
med fine sand	No	% medium and fine in the sediment
mud	No	% mud in the sediment

21.3 Data exploration

Figure 21.2 shows the pairplot for the explanatory variables gravel, large sand, medium fine sand and mud content. The nearly straight, close to 45-degree, straight lines in this plot indicate a strong linear relationship among mud, medium fine sand and large sand content of the sediment. Using all three explanatory variables would lead to serious problems with forward and backward selection procedures in GLM or GAM. Presenting a model that includes all these variables is not a good solution either. The reason for this is that it does not make sense to use a model where some of the explanatory variables mirror the same information. The strong linearity among these three factors allows two to be dropped from the analysis. This reduces the number of parameters in the analysis, and generally the fewer parameters the better, as long as you do not lose too much information. In this instance, we omitted the medium fine sand and the large sand content, using only the mud content and the gravel content for analysis. Note this is an ecological choice rather than a statistical one.

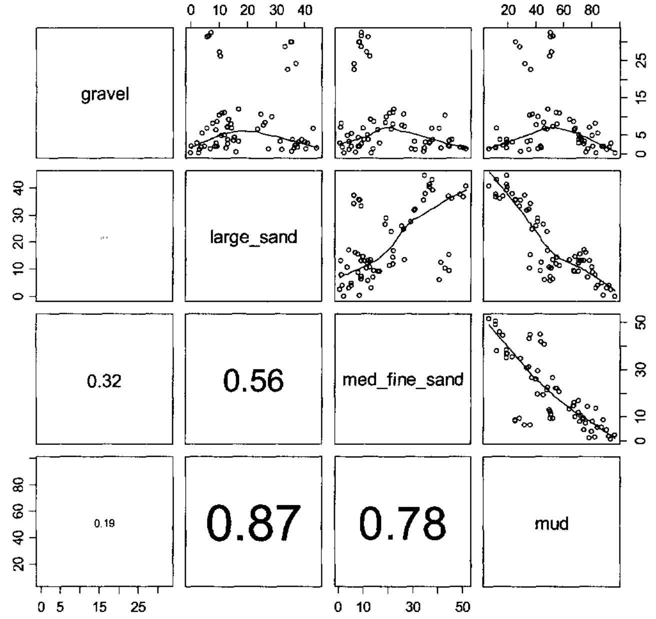

Figure 21.2. Pairplot of four explanatory variables indicating collinearity. The number below the diagonal are (absolute) correlations coefficients. The font size of the cross-correlation is proportional to its strength. The lines in the upper diagonal panels are LOESS smoothers.

A pairplot (Figure 21.3) of the remaining continuous variables gives no immediate indications of further collinearity between the explanatory variables. The first row of graphs shows the relationship between *S. solea* and each of the continuous explanatory variables. Some of the shapes (e.g., for depth and salinity) show the typical fit of a logistic generalised linear model (Chapter 6), and this is a useful guide for choosing an appropriate analysis. Looking at possible relationships between the response variable and nominal variables, a design plot (Figure 21.4) indicates that the mean value in the first season (labelled as 0) is higher than in the second season (labelled as 1), month nine has a considerably lower mean value, and the mean value in area one is nearly twice as high as in the other areas. As season is just a coarser version of month, both these variables should not be used in the same GAM or GLM model (collinearity). It was decided to use month instead of season as it provides more information.

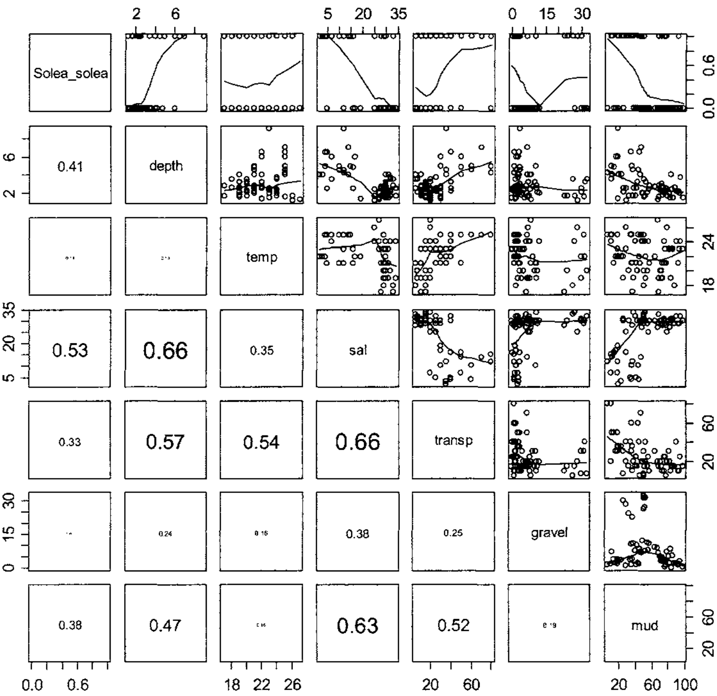

Figure 21.3. Pairplot of *S. solea* and selected explanatory variables. The lower diagonal panels contain the (absolute) correlations. The font size is proportional to the value. The upper diagonal panels show the pair-wise scatterplots. A smoothing line was added.

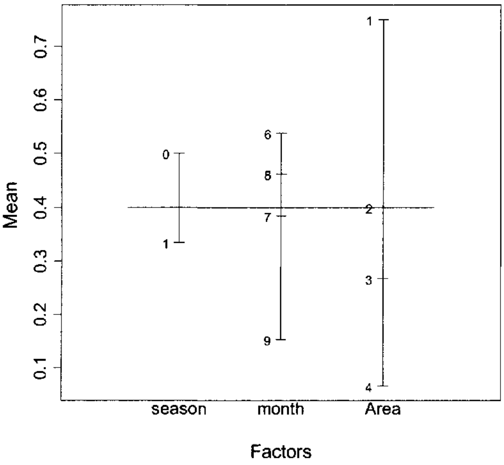

Figure 21.4. Design plot for *S. solea* presence–absence data. The *y*-axis shows mean values per class of the nominal variable. Highest mean values are in month 6 and in area 1.

Figure 21.5 shows a coplot of *S. solea* versus salinity conditional on the nominal variable month. We deliberately selected month as a non-nominal variable in the coplot. As a result, the software groups data from different months and makes a scatterplot between *S. solea* and salinity for those months. For example, the lower left panel in Figure 21.5 shows a scatterplot of *S. solea* and salinity for the samples measured in the months 5 and 6. The lower right visualises the same, expect for samples from months 6 and 7. All panels show a similar *S. solea*-salinity relationship, except for the panel that includes samples from month 9. This is another indication (month 9 also had a particularly low mean in the design plot) that we need to give special attention to month 9 in the GLM and GAM models. A coplot in which month was used as a nominal variable resulted in five panels with less data per panel.

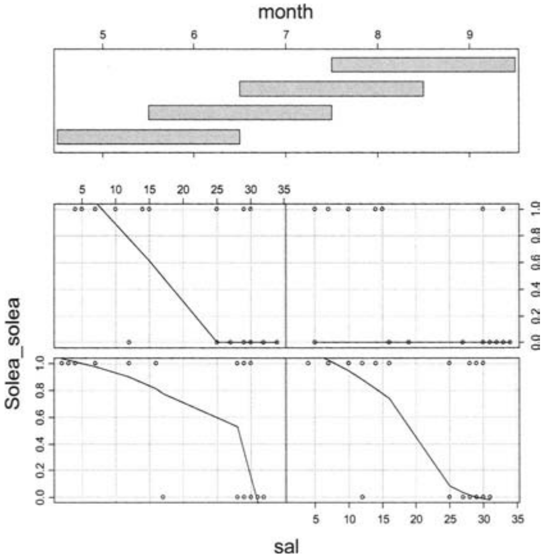

Figure 21.5. Coplot between *S. solea* and salinity, conditional on month. A smoother was fitted in each panel. The lower left panel contain the data from months 5 and 6, the lower right from months 6 and 7, etc. A smoothing curve was added.

21.4 Classification trees

Even though the above exploratory tools allow some variables to be dropped from the analysis, the number of explanatory variables is still relatively large. A classification tree allows a more detailed investigation into the relative importance of these remaining variables. The classification tree (Figure 21.6) indicates that salinity is the most important explanatory variable. A high probability of finding *S. solea* is obtained for samples with salinity smaller than 15.5 and a gravel content larger than 1.34. Using a pruning diagram (Figure 21.7) indicates that the tree presented in Figure 21.6 is sub-optimal (Chapter 8), and that a tree of size two would be optimal. Due to the small sample size, it may be wise to reapply the cross-validation procedure several times using different starting values, and this indicated that one should either use a tree of size two or six. Selecting which one is best is subjective. However, all indicate that salinity is the most important variable, with the importance of gravel, and the variables further down the tree open to argument.

As the same variable appears more than once in the tree, it indicates a weak non-linear relationship between the response variable and the explanatory variables. This suggests that the next step should be a GAM (Chapter 7) as this can deal with non-linearity, and the results from the tree suggest only one important variable to be identified by the GAM (the optimal tree size is 2).

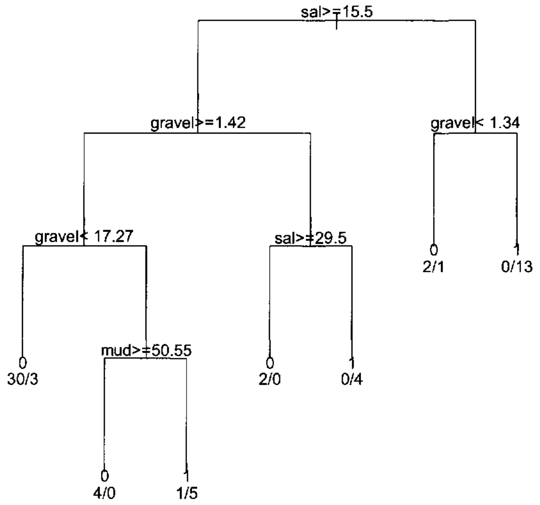

Figure 21.6. Classification tree for *S. solea*. If a statement is true, follow the left branch. Numbers at the end of a branch are the predicted group (1 = presence, 0 = absence) and the classifications per group.

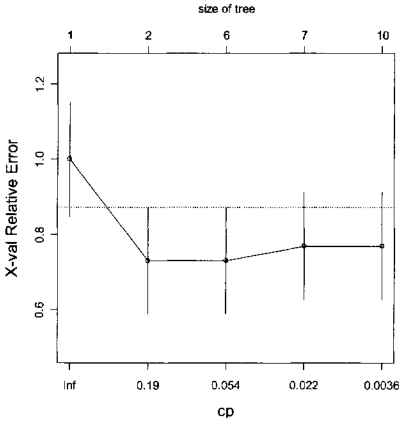

Figure 21.7. Results of cross-validation. The 1-SE rule dictates to select the left-most tree for which the mean relative error is below the dotted line, in this case a tree of size 2.

21.5 Generalised additive modelling

A GAM using the binomial distribution and logistic link function (Chapter 7) was used to relate *S. solea* presence–absence data and the explanatory variables. Strictly speaking, the model is not a GAM with a binomial distribution but rather one with a Bernoulli distribution. With Bernoulli models overdispersion cannot occur. To find the optimal set of explanatory variables, a forward selection was applied. Table 21.2 gives the AIC using each of the continuous explanatory variables as a single explanatory variable. The lower the AIC value, the better it is, and cross-validation was used to estimate the optimal degrees of freedom for each smoother. The GAM identified salinity as the best single explanatory variable. In the next step of the forward selection procedure, two explanatory variables were used, one of them salinity. None of the combinations led to a model with a smaller AIC value (compared with the one with salinity) with significant smoothers. These results indicate that the GAM model using only salinity is the best model. The effect of salinity is presented in Figure 21.8. The dotted lines represent a 95% confidence interval. The cross-validation estimated 1 degree of freedom for the smoother. This means that a GLM should be applied instead of a GAM (GAM with a smoother with 1 degrees of freedom is equal to a GLM). For reasons of completeness, the numerical output for the GAM is given below.

```
Parametric coefficients:
            Estimate     std. err.      t-ratio        p-value
Intercept   -0.42        0.29           -1.40          0.16

Approximate significance of smooth terms:
            edf    chi.sq       p-value
s(sal)      1       13.823      <0.001

R-sq.(adj)=0.26, deviance explained = 21.6%, n = 65, Deviance = 68.56
```

Table 21.2. AIC values for first step in the forward selection procedure. Cross-validation was applied in each step. A Binomial GAM was used. The row in bold typeface is the most optimal model.

Single explanatory variable in GAM	AIC	edf
Depth	80.14	1
Salinity	**72.56**	**1**
Temperature	90.63	1
Transportation	83.92	2.7
Gravel	85.35	3.5
Mud	80.4	4.1

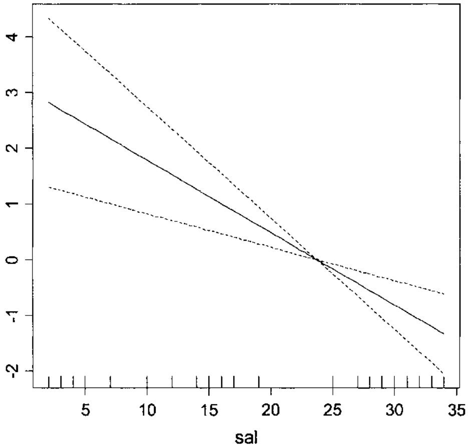

Figure 21.8. Partial fit of salinity. The x-axis shows the salinity gradient and the y-axis is the contribution of the smoothing function f(salinity) in the model logit(Y) = intercept + f(salinity).

21.6 Generalised linear modelling

The advantage of GLM over GAM is that GLM is parametric. Using the same explanatory variables as in the GAM, a backward selection, and a combination of a forward and backward selection was applied. Both approaches gave a model where salinity, month, temperature and gravel (in order of importance) were selected. The estimated parameters, standard errors, z-values and p-values are given below. Note that salinity is highly significant. Other parameters are weakly significant. Instead of assessing the importance of individual explanatory variables using the z-values (or t-values) in Table 21.3, we drop one term in turn and compare deviances of the full and nested model with each other using the Chi-square (Chapter 6). The output of this approach is given in Table 21.4 and shows that all variables are significant at the 5% level, except for gravel.

Table 21.3. Estimated parameters obtained by the GLM model containing temperature, salinity, gravel and month. The AIC is 70.145.

	Estimate	Std. Error	*t*-value	*p*-value
(Intercept)	23.35	9.69	2.41	0.16
Temperature	−0.89	0.42	−2.09	0.03
Salinity	−0.26	0.07	−3.78	0.00
Gravel	0.06	0.03	1.68	0.09
factor(month)6	2.01	1.21	1.66	0.09
factor(month)7	3.37	2.14	1.57	0.11
factor(month)8	3.86	2.01	1.92	0.05
factor(month)9	−2.20	1.32	−1.66	0.09

Table 21.4. Change in deviance and corresponding Chi-square values if one variable is dropped from the model.

Variable to be dropped	df	Deviance	AIC	Chi-square value	*p*-value
<none>		54.14	70.14		
Temperature	1	59.82	74.22	5.68	0.01
Salinity	1	81.33	97.25	27.19	<0.001
Gravel	1	57.08	71.28	2.93	0.08
Factor(Month)	4	65.88	74.71	11.73	0.01

The partial fits (Chapter 5) in Figure 21.9 show the contribution of the individual explanatory variables, while taking into account of the other variables in the model. One of the reasons that gravel is included in the model might be due to the isolated set of samples for high gravel values. Month shows a weak seasonal pattern, with month 9 clearly displaying the lowest values. The variables salinity and temperature have a negative relationship with *S. solea* and a positive relationship with gravel. A detailed model validation included Cook distance values, histograms and QQ-plots of residuals, hat values, changes in fit and parameters after leaving out one variable, and did not indicate any serious problems (Chapter 5).

We now have the output from three different techniques: classification trees, GAM and GLM. All three techniques indicate that salinity is significant, and the most important variable. Gravel was also selected by all techniques, but the *p*-values for this variable indicated a very weak relationship.

Based on the results of the GLM, we re-applied the GAM model using salinity, temperature, gravel and month as explanatory variables. The reason for this is that if the GAM curves obtained by this analysis are straight lines (or can be considered as approximately straight lines within the point-wise confidence bands), then we have a confirmation that the GLM model is indeed the most appropriate method. We used cross-validation in the GAM model to find the optimal degrees of freedom. The output (given below) shows that temperature and salinity are fitted as a linear component, but gravel has a modest non-linear effect.

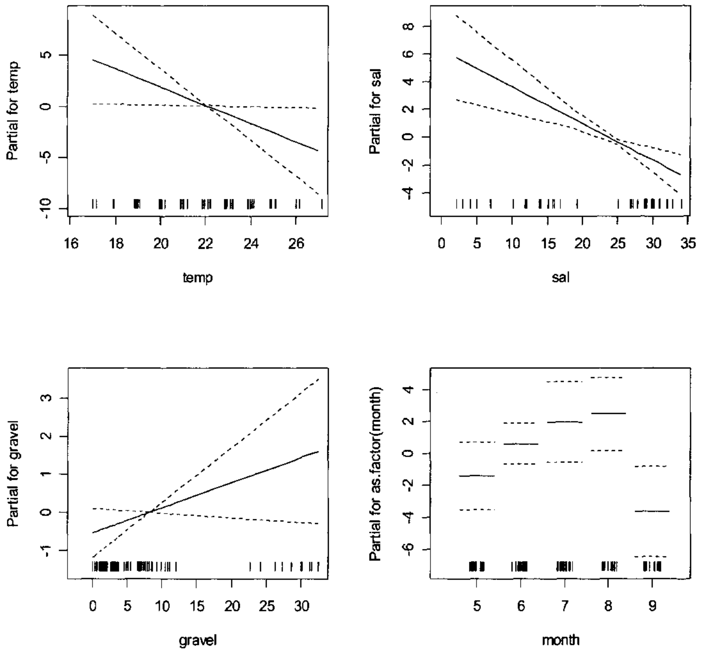

Figure 21.9. Partial fits of GLM model.

```
Parametric coefficients:
                    Estimate   std. err.   t-ratio      p-value
    (Intercept)      -1.71      1.30       -1.38        0.16
factor(month)6        1.72      1.30        1.32        0.18
factor(month)7        2.68      2.26        1.18        0.23
factor(month)8        3.37      2.15        1.56        0.11
factor(month)9       -2.57      1.44       -1.78        0.07

Approximate significance of smooth terms:
              edf       chi.sq          p-value
s(temp)       1.11      6.45            0.37
s(sal)        1.07      15.96           0.02
s(gravel)     3.25      10.81           0.28

R-sq.(adj) =  0.42   Deviance explained = 46.8%
n = 65, Deviance = 46.58 , AIC = 67.47
```

Although the residuals did not show any patterns (which is good), the confidence intervals around gravel (not shown here) were rather large. The *p*-values for the continuous smoothers indicate a linear salinity effect, but no temperature or gravel effect.

21.7 Discussion

An outline of the analysis workflow is shown in Figure 21.10. The data exploration suggested omitting two continuous and one nominal explanatory variable from the analysis to avoid collinearity problems in the GLM and GAM models. The variables omitted were medium fine sand and large sand content of the sediment and season. Once the optimal models were fitted using the remaining variables, these variables were added back into the analysis as a check, but they did not improve the model.

Classification models were applied next, indicating a strong salinity effect, and possible non-linear relationships, as the same variables occurred at different branches. At this point, you could decide to apply a GLM, but we decided to apply a GAM first as this would help to visualise the type of relationships between *S. solea* and the explanatory variables. The GAM results indicated that the main variable, salinity, was having a linear effect, and therefore we continued with a GLM. A detailed forward and backward selection process in the GLM indicated a month (nominal), temperature, salinity and gravel effect. However, as a GLM is imposing linear relationships (on the predictor scale; Chapter 6), we decided to verify the optimal GLM with a GAM. If the GAM, using month, temperature, and gravel, indicates that the relationships are indeed linear, then we can be confident that the GLM results are correct. We can then present them as our final results. Both temperature and salinity were estimated as linear components by the GAM, but gravel was slightly non-linear. However, in such a model (not presented here) temperature and gravel were not significant.

All the techniques we used showed a significant and strong salinity effect. As to the ecological interpretation, *S. solea* seems to prefer relatively low salinities, and there is a higher probability of catching *S. solea* in months 7 and 8, than in any of the other months.

Obviously, it is difficult to say whether there is a cause-and-effect relationship between salinity and the probability of *S. solea* occurrence. There might be other variables, not measured, but still highly correlated with salinity that are the real driving factors. Despite these difficulties in identifying a cause–effect relationships, the influence of salinity on the abundance of *S. solea* has been reported by other authors. Riley et al. (1981), for the UK waters, and Marchand and Masson (1988) for waters off North France, where it was found that sole less than one year old prefer salinities between 10% and 33%.

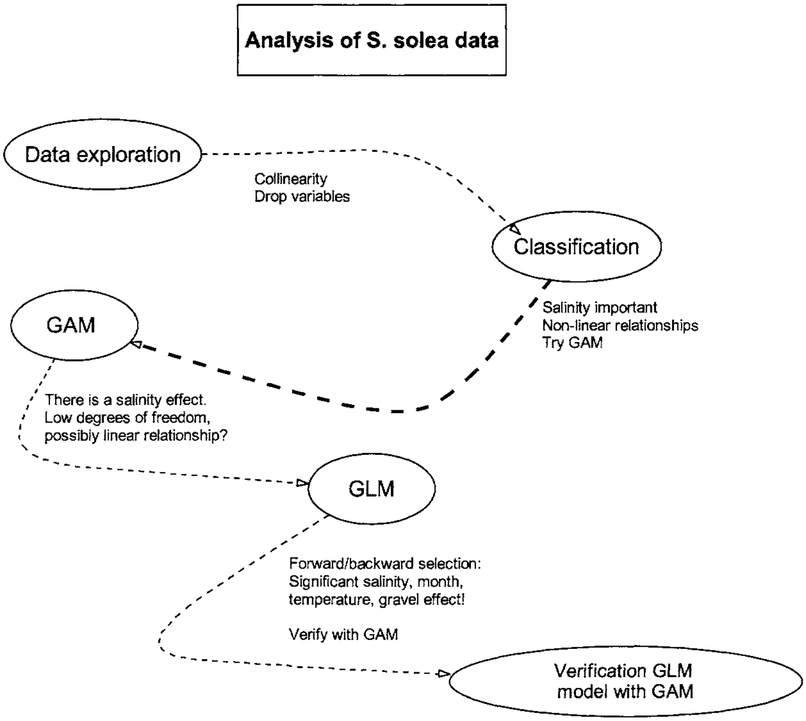

Figure 21.10. Outline of data analysis approach.

The use of several techniques applied to the same dataset allowed us to evaluate the influence on the results due to the selection of a particular statistical technique. Although it is difficult to identify a single best method for comparing different techniques, applied to the same dataset, the approach adopted here should give a useful starting point. The fact that the GLM, GAM and tree models find different optimal models shows the danger of forward selection and relying on only one statistical technique. Even within a particular method, e.g., GAM, one might find different results depending on the model selection strategy, e.g., the use of cross-validation, AIC, or Chi-square deviance tests. However the use of GLM and GAM methods allow a fine-tuning procedure that can lead to marked improvements in the models considered.

Acknowledgement

We would like to thank Ian Tuck and Graham Pierce for useful comments on an earlier draft.