Developing Essential Fish Habitat maps for fish and shellfish species in Scotland
Annex 2. Decision Tree Models

Author: Anita Franco

This technical annex provides details on the data-based modelling approach used in the study and guidance on how to interpret and apply them for predicting the potential distribution of aggregations of a species/life stage.

A2.1 Decision tree modelling approach

Decision tree models, specifically Classification trees, were calibrated to predict the presence/absence of aggregations of a species/life stage based on the associated environmental conditions.

Classification trees are a type of supervised learning algorithm belonging to the CART family (Classification and Regression Trees) that allows the analysis of the relationship between one response variable (a categorical one in the specific case of classification trees) and several explanatory variables (Guisan et al. 2000, Faraway 2006, Zuur et al. 2007). They organise explanatory variables in a hierarchical way, based on their effect on the response variable, thus resulting in a tree-like structure (see section AX.3 for the structure and interpretation of decision trees).

Compared to other classification approaches, decision trees closely mirror human decision-making and have several advantages (De'ath and Fabricius 2000, Zuur et al., 2007). The hierarchical nature of the trees allows to identify the relative importance of different explanatory variables. They are similar to additive models in that they represent a compromise between the linear model and the completely nonparametric approach. In fact, they can accommodate the non-linearity and interaction between explanatory variables, as well as missing values (which may often been present for environmental variables in the datasets).

A major advantage of decision trees is also that they are intuitively very easy to understand, and the resulting algorithm (i.e., the combination of environmental ranges that can be used to predict the occurrence of aggregations of a certain life stage) can be easily applied to a new environmental scenario to obtain predictions. As a result, they can be more easily communicated to a non-expert audience and used even without any technical expertise or statistical package needed.

The model is calibrated by adjusting the selected mathematical model for the specific data on which the model is trained on (Guisan et al. 2000). In fact, the model does not need to be specified at the outset, as the true starting point is the algorithm created during the construction of the tree (Faraway 2006).

A2.2 Modelling protocol

All analyses were conducted using R version 3.3.3 (R Core Team 2017) and the R-package "rpart" (Therneau et al. 2019).

Fish survey and environmental data were combined into individual datasets for specific species/life stage, which including the presence/absence of aggregations (response variable) at individual survey events (e.g. haul) and the associated environmental conditions (explanatory variables) as extracted from environmental data layers. Each dataset was randomly divided into a train subset, including 80% of the data used for calibrating the model, and a test subset including the remaining data (20%) for model statistical validation (see assessment of model performance in Annex 3)

Before modelling, a preliminary analysis of collinearity was undertaken to exclude redundant/collinear environmental variables. Collinearity occurs when there are high correlations among predictor variables, leading to unreliable and unstable estimates of the model parameters. Classification tree analysis only allows one of any set of correlated variables to enter the model at any given split (the variable that best classifies the data is selected). As the data are split into smaller groups throughout the modelling process, the relationships among explanatory variables may change, and this may affect the model results, thus justifying exclusion of highly correlated predictors (Lawler and Edwards 2002). A Variance Inflation Factor (VIF) <10 and/or Pearson's correlation coefficient ≥0.8 were used as indicators of collinearity and the environmental variables were excluded from the analysis accordingly (Zuur et al. 2009). As a result, the following variables were excluded from the individual datasets:

- Lesser sandeel: Depth (positively correlated with MLT); VegH01 was also excluded as all values were 0 in the dataset;
- Anglerfish (juveniles): Substratum type (high VIF);
- Cod, Haddock, Norway pout and Whiting (spawning): MLT (positively correlated with Depth).

A model was calibrated on each train dataset by initially including all the environmental variables as potential predictors. Automatic selection procedures within the model algorithm allowed to identify the initial best model, i.e. based on a subset of explanatory variables that identify the parameters that best explain the variability in the fish data.

Further model selection was undertaken through a combination of cross-validation (estimating the predictive ability of the model) and truncating in order to reduce the tree to a more 'optimal' number of terminal nodes as large tree can cause overfitting and may be more difficult to interpret (Faraway 2006). Specifically, "pruning" of the tree model was undertaken by applying the "one standard deviation (1-SE)" rule through examination of the Cp plot. Cp (the complexity parameter) controls the size of the tree (the latter decreases with increasing Cp). Pruning the tree (hence using a more simplified tree and higher Cp) usually increases the cross-validation prediction error, as shown in the Cp plot. An

acceptable compromise must be found between the level of pruning and the model error. The parsimonious “one standard deviation (1-SE)” rule allows the tree to be pruned to a maximum Cp value for which the resulting tree has a rate of error within one standard deviation of the minimum error (Faraway 2006, Zuur et al. 2007). Therefore, this rule allows simplifying the tree while maintaining an acceptable degree of error.

Ten iterations of the Cp plot were undertaken and all the resulting models meeting the 1-SE rule were initially selected. The final best model was selected as the model that provided the best results on validation, i.e. the model with the best performance (highest F1 score; see assessment of model performance in Annex 3).

A2.3 How to interpret and use decision trees

A decision tree model is visually represented according to the tree structure in Figure A2.1.

The decision tree starts from a root node that represents the full population of observations. This is divided through a series of splits into pairs of homogenous subsets of observations (sub-nodes) along a branch (i.e. a sub-section of the tree). Where a sub-node splits into further sub-nodes, this is called decision node. A node at the end of a branch and that does not split further is called terminal node (or leaf).

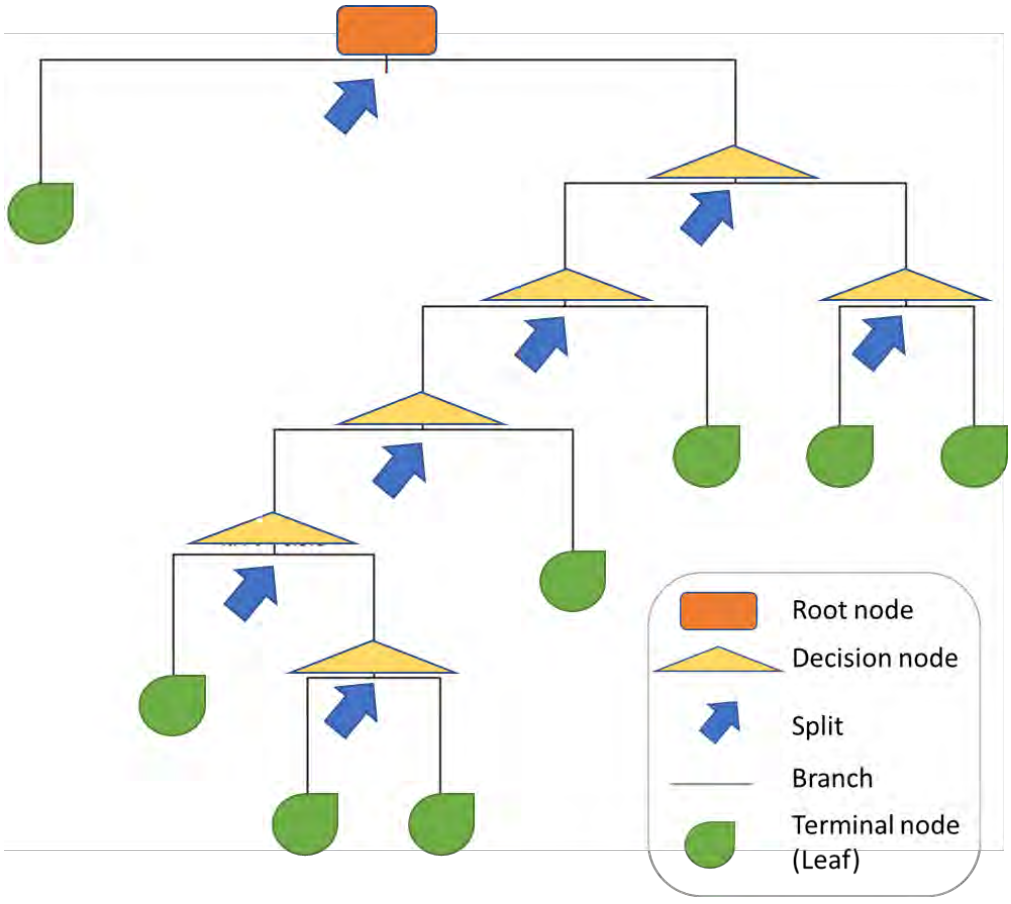

Figure A2.1. Structure of a decision tree and its key elements.

In a Classification tree such as the one calibrated in this study, the root node is the full population of observations as classes of presence/absence of aggregations. Recursive binary splitting is applied at the root node and then at each of the decision nodes to grow the tree. Each split is defined based on alternative conditions for a specific environmental variable (e.g. < or ≥ of a certain value). The value indicated at the split on the tree (see for example classification tree for Plaice juvenile aggregations in Figure A2.2) represents the condition predicted along the branch on the left-hand side of the node (e.g. MLT ≥18.23 at the first split in Figure A2.2), whereas the alternative condition (e.g. MLT <18.23) is associated with the right-hand branch. As a result of the recursive splitting, a hierarchy can be established between the environmental predictors in the tree model, with those at the top being more important in affecting the response variable. Where the same predictor determines multiple splits along a branch, this indicates a non-linear effect of that variable on the prediction.

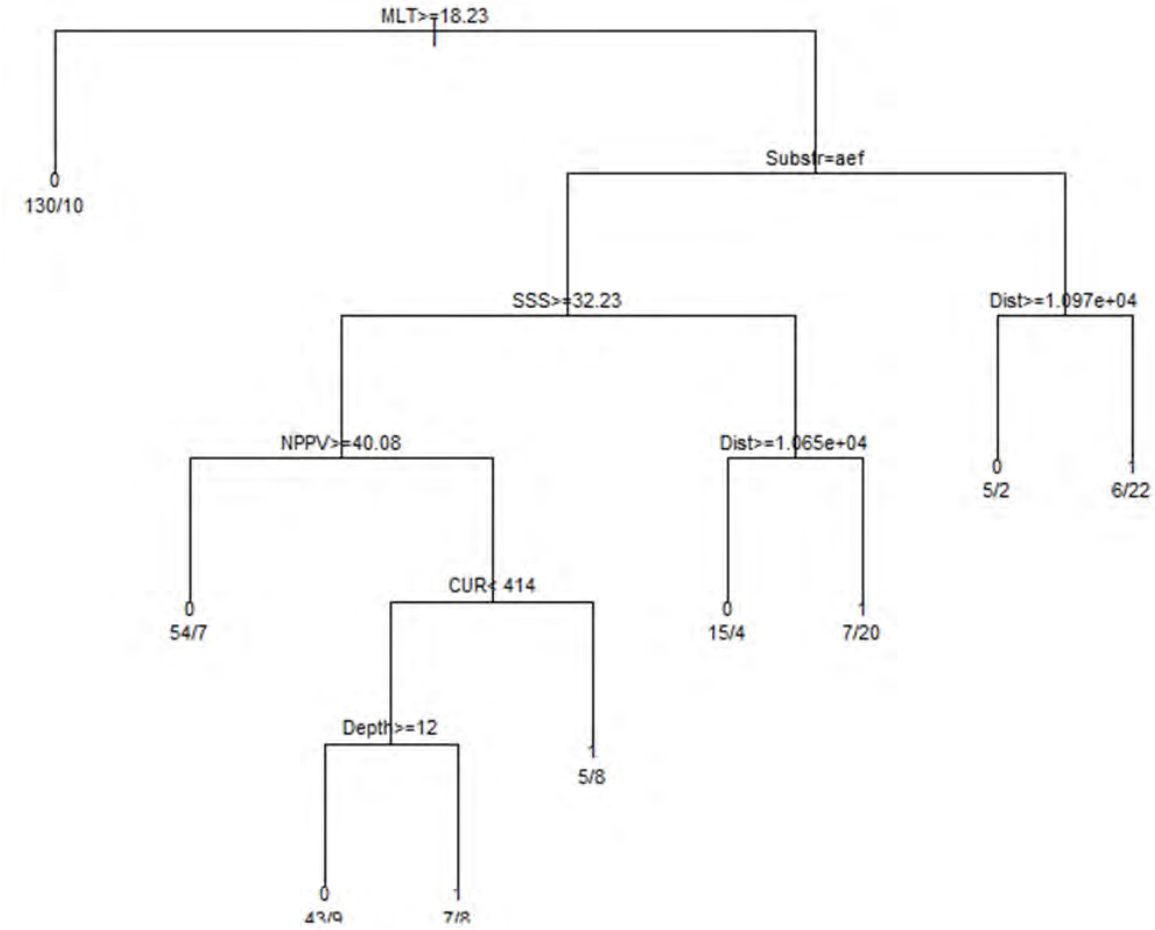

Figure A2.2. Example of classification tree obtained for Plaice juvenile aggregations based on selected environmental predictors (MLT, thickness of the mixed layer (m); Dist, distance from the coast (m); Depth, water depth (m); SSS, sea surface salinity; CUR, current energy at the seabed (N/m^2); NPPV, net primary production (carbon per unit volume of seawater); Substr, substrate type classified as a = Coarse substrate, b = Fine mud, c = Mixed sediment, d = Muddy sand, e = Sand, f = Sandy mud, and g = Sandy mud or Muddy sand).

The resulting predicted class (presence as 1 or absence 0) is shown at the leaf end of a branch (terminal node) and the combination of environmental criteria leading to that prediction is given by the conditions set along the branch. At the leaf prediction, the proportion of class observations in the training dataset that originate the leaf is also indicated (as number of absences/number of presences). This can be used to assess the probability of presence or absence predicted by the leaf.

For example, the model in Figure A2.2 predicts absence of summer aggregations of juvenile Plaice where MLT ≥18.23 m, with 130 absence and 10 presence training observations included in that leaf prediction, suggesting that the absence class is predicted under those environmental conditions with 93% of success (i.e. 0.93 probability of absence =130/(130+10)), whereas the probability of presence of aggregations predicted in those conditions is 7% (10/(130+10) = 0.07).

The same model predicts the presence of aggregations where MLT <18.23 m, the Substrate is Fine mud, Mixed sediment, Muddy sand or Sandy mud/ Muddy sand and the Distance from the coast is <10,970 m. Where all these conditions are satisfied, the presence is predicted with a 79% probability (22/(6+22)=0.79) while 21% of the observations were incorrectly classified by that leaf prediction.

As each leaf prediction (as predicted class, presence or absence, and further qualified by the probability of the prediction) can be associated with a specific set of environmental criteria clearly defined by the model, its application does not require special expertise of statistical tools. Rather, it only requires that the environmental variables used as model predictors are estimated under the desired new scenario (a new area to be assessed, or an area where environmental conditions have changed, e.g. over the years, or are expected to change, e.g. due to climate change). Where the specific environmental criteria defined along different branches of the decision tree are met in the new scenario, the correspondent prediction of whether the habitat is potentially suitable to support aggregations of the species/life stage, and with which probability can be obtained.

The model predictions (and the associated environmental criteria defined in the decision tree) need to be put into the context of the environmental variability represented in the data that were used to create the model (the environmental ranges associated with the survey data). For example, only sedimentary substrata were covered by the surveys that originated the model for juvenile Plaice in Figure A2.2, and the survey locations closest to the coast were at a distance of 744 m. Therefore, the model predictive ability is limited by these ranges, and any predictions outside them (e.g. for areas of hard substrata or closer to the coast) are not considered to be reliable.

A2.4 References

De'ath, G. and Fabricius, K.E. (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology, 81(11): 3178–3192.

Faraway, J.J. (2006) Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. Chapman and Hall/CRC, 312 pp.

Guisan, A., Edwards Jr., T.C. and Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling, 157: 89–100.

Lawler, J.J. and Edwards, Jr. T. (2002) Landscape patterns as habitat predictors: building and testing models for cavity-nesting birds in the Uinta Mountains of Utah, USA. Landscape Ecology, 17: 233–245.

R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Therneau, T., Atkinson, B. and Ripley, B. (2019) Package 'rpart': Recursive Partitioning and Regression Trees. R package version 4.1-16. https://cran.r-project.org/package=rpart

Zuur, A.F., Ieno, E.N. and Smith, G.M., Eds. (2007) Chapter 9. Univariate tree models. In: Analysing Ecological Data. Springer: 143-161.

Zuur, A.F., Ieno, E.N. and Elphick, C.S. (2009) A protocol for data exploration to avoid common statistical problems. Methods in Ecology & Evolution, 1: 3-14.