Developing Essential Fish Habitat maps for fish and shellfish species in Scotland
Annex 3. Confidence assessment of data-based models

Author: Anita Franco

Confidence is a key element associated with models and their predictions, as it provides guidance about how much the results can be trusted. This technical annex gives details on the methodology used to estimate confidence for the models and associated spatial predictions presented in the main report.

A3.1 Methodology

A3.1.1 Overall confidence

The overall confidence associated with the spatial output indicating the potential location of a species' EFH was assessed by combining different elements of confidence associated with the modelling and mapping process, as described below.

Model performance (overall)

The statistical validation of a classification model consists on the application of the model to a test dataset (20% of the survey data were used in our case), and the comparison of the model prediction against the true observation in the dataset in order to assess the model performance. A confusion matrix is used, which, for a binary classification such as the presence/absence in this study, is as in Figure A3.1.

Confusion matrix		Model prediction	
		Absence	Presence
True observation	Absence	TN	FP
	Presence	FN	TP

Figure A3.1. Confusion matrix for presence v. absence classification. TN true negative, TP true positive, FP false positive, FN false negative (the sum of these four values is the total number of observations in the dataset where the model is predicted).

A typical measure of confidence in a classification model is its classification accuracy, i.e. the number of correct predictions from all predictions made (calculated as TP+TN divided by the sum of all elements in the confusion matrix). However, this may not be an adequate performance measure in cases where the number of 'negative' observations (absences) is much greater than the number of positive observations (presence) (Kubat et al. 1998), as in the datasets analysed in this study (where 25% is presence, given the definition of aggregations used). With such large class imbalance, the model can predict the value of the majority class (absence) for all predictions (even for true presence observations) and achieve a high classification accuracy (this is called the accuracy paradox). For example, in a dataset with 25 presence observations and 75 absence observations, a model that predicts all 100 observations as absence (hence TN = 75, FN = 25 and TP = FP = 0) would have an accuracy of 75%, even if all the actual presence observations are not correctly predicted by the model. Given that our aim is to calibrate a model that is able to predict where aggregations of a species/life stage may potentially occur, the ability of predicting presence is key to define the model performance, and therefore the accuracy estimate is not suitable in this case.

The F1 score (Lewis and Gale 1994) was used to estimate the model performance (C_m) in predicting the presence/absence of aggregations (PAaggr) of a species' life stage (considered as an indicator of EFH). This is a metric which takes into account both model precision and sensitivity (the latter also known as recall or true positive rate), as per the equations below:

$$C_m = F1\ Score = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity} \quad \text{(Eq. 1),}$$

$$Precision = \frac{TP}{TP+FP} \quad \text{(Eq. 2),}$$

$$Sensitivity = \frac{TP}{TP+FN} \text{(Eq. 3),}$$

where TP, FP and FN are respectively True Positive, False Positive and False Negative values as obtained from the confusion matrix. F1 score can range from 0 to 1, with 1 representing a model that perfectly classifies each observation into the correct class and 0 representing a model that is unable to classify any observation into the correct class. This metric has been indicated as a better estimate of model performance compared to other metrics in the case of imbalanced datasets and/or when the interest is mostly about the positive class (Czakan 2021).

Fish survey data (overall)

The confidence associated with the fish survey data (C_f) on which the model was calibrated was assessed based on the ability of the survey to reliably represent the distribution of the species life stage of interest. Different aspects of the survey data were considered in the assessment, such as:

- Efficiency of the sampling method in sampling the species life stage;
- Timeliness of sampling design, accounting for both coverage of the season(s) most relevant to the species life stage, and of the years within the selected study period (2010-2020);
- Spatial confidence of sampling design, accounting for both geographical coverage of UK waters and the likelihood of coverage of the EFH of interest (considering

knowledge of them from literature and the likely habitats that can be sampled with the survey method);

- Confidence in the identified life stage as indicator of EFH, reflecting how well the life stage, as identified in the samples, is likely to represent the distribution of EFH.

Each of the above elements was scored (1 to 5, for low to high confidence) and the total mean score was calculated across elements. The final score was reproportioned to a maximum of 1 for inclusion in the overall confidence calculations.

Environmental data (overall)

The confidence associated with the environmental data (C_e) as extracted from spatial layers and used for the model calibration was assessed. Similar criteria as for the survey data were used for this assessment, including methodology and quality standards, timeliness (for non-persistent variables only) and spatial confidence. Information available from confidence/quality assessments provided in documents associated with the data layers was used for this assessment. Where confidence maps were available for the environmental variable from the data layers, the mean confidence for the overall map was considered as an overall assessment for that variable. The final (mean) confidence score attributed to each environmental variable was rescaled to a maximum of 1 for inclusion in the overall confidence calculations.

For each species model, the overall confidence of the environmental data used in the model was calculated as follows:

$$C_e = \frac{\sum_i (E_i * I_i)}{\sum_i I_i} \quad \text{(Eq. 4),}$$

Where E_i is the mean confidence of the individual (i-th) environmental data input, and I_i is the score attributed to each environmental variable based on its importance in determining the model (as obtained from summary statistics of the model). Only the set of variables actually included in the model as predictors were considered.

Total confidence (overall)

The overall total confidence associated to a model spatial output (C) was calculated as follows:

$$C = C_m * \frac{C_f + C_e}{2} \qquad \text{(Eq. 5).}$$

This equates to weighting the model performance score (C_m) by the mean confidence score of the input data (C_f and C_e).

The resulting score (0-1 scale) was categorised as follows, and colour coded accordingly:

- High confidence: C ≥0.8 (Blue);
- Good confidence: C ≥0.6 and <0.8 (Green);
- Moderate confidence: C ≥0.4 and <0.6 (Yellow);

- Poor confidence: C ≥0.2 and <0.4 (Amber);
- Low confidence: C <0.2 (Red);

A3.1.2 Spatial confidence

It is acknowledged that the confidence may not be homogenously distributed across the predicted space. This spatial variability may be due to variability in the predictive error associated with the model prediction under a specific set of environmental conditions (i.e. 'leaf' prediction in the classification tree) or the variability in quality and precision of the environmental variable estimates used for the model spatial prediction. Therefore, confidence was also assessed spatially (i.e. for each modelled grid cell), using a similar approach as used for the overall confidence, with some modifications as described below.

Model performance (spatial)

In this case the confidence associated with the specific model prediction within a grid cell was calculated as:

$$Grid\ cell\ C_m = (ClassProb - 0.5) * 2 \qquad \text{(Eq. 6)},$$

where *ClassProb* is the probability of the predicted class (presence or absence) in the grid cell (also referred to as leaf prediction as it is the prediction resulting from one branch of the classification tree). *ClassProb* varies between 0.5 (i.e. when the class is predicted with 50% probability, i.e. both classes have same chance of occurring, and therefore the class allocation for the leaf prediction is considered random) and 1 (i.e. when all observations in the leaf were correctly predicted by the model). Equation 6 rescales the probability value, so that a confidence value of 0 is allocated to the former case, and 1 to the latter.

Fish survey data (spatial)

As no spatial variability was identified for the confidence associated with the fish survey data (C_f), this element was not included in the spatial confidence assessment.

Environmental data (spatial)

The confidence associated with the individual environmental data at grid cell level was calculated by considering both (i) the quality of the spatial estimates, as indicated by confidence maps associated with the individual data layers (where available), and (ii) the precision associated with the use of a mean estimate of a variable allocated to each grid cell, as derived from the coefficient of variation (CV) calculated during data processing. These two levels of spatial confidence and their integration in a final confidence estimate are described below in detail.

Quality of the spatial estimates

The source for this part of the confidence assessment were the confidence maps associated with the source environmental data layers (where provided) and accounting for spatial variability in the confidence of the associated environmental variable.

Confidence values (ranging 0-3) were provided with spatial estimates of kinetic energy at the seabed associated with current and waves in the source layers. These spatial values were assigned to individual grid cells as a spatial confidence estimate for CUR and WAV variables. The values were standardised to a 0-1 range for integration into the final confidence assessment.

The EMODnet Bathymetry layer included an assessment of quality of the bathymetric product. This was expressed as a Combined quality index (CQI, as %) associated with the bathymetry estimates, combining separate assessments of the accuracy of the survey, temporal representativity/consistency, completeness, and age of the survey originating the bathymetric data. The mean CQI for each grid cell (weighted by the area covered in the cell by data with different CQI) was used as an estimate of spatial confidence associated to this type of data. This informed the spatial confidence assessment of both Depth and Slope variables.

Confidence values (as 1-3 score) are provided for substratum type classes in the EMODnet Substrate environmental layer[1]. The predominant confidence associated with the dominant substratum type identified for each grid cell was used as an estimate of spatial confidence.

No confidence maps were available for the other environmental layers.

Precision of the spatial estimates

Where, for the purpose of predicting the model and mapping the result, an environmental variable was estimated from a data layer as the mean value within a cell, an additional confidence estimate was derived to account for precision of such estimate (i.e. the spatial and/or temporal variability of the environmental values around the estimated mean). This was expressed as follows:

$$E_{ij(precision)} = 1 - \frac{CVij}{max(CV)} \quad \text{(Eq. 7),}$$

where $E_{ij(precision)}$ is the confidence of the individual (i-th) environmental variable in the j-th grid cell, CV_{ij} is the Coefficient of Variation[2] of the data for that variable within the grid cell, and max(CV) is the maximum CV recorded across all variables in the grid. CV accounts for the variability of the data over space alone (within a grid cell, for Depth, CUR, WAV) and space and time (across all months within the relevant seasons and across years within the study period 2010-20, for temporally variant variables, i.e. MLT, NPPV, SBT, SST, SSS). Equation 7 standardises confidence within the 0-1 value range so that lower confidence

[1] For substrate classification of the grid cells, INFOMAR data layer was also used to integrate missing/unclassified areas in EMODnet. However, INFOMAR data only accounted for 0.3% of the identified substrate types in the grid, the remaining 99.7% of allocated substrate types being derived from EMODnet. Therefore, using EMODnet confidence alone was considered to be sufficient for a realistic estimate of confidence of the substrate data used.

[2] CV = Standard Deviation / Mean

values (closer to 0) are associated with grid cells where higher data variability around the mean estimate (i.e. lower precision) was observed.

Spatial confidence of the environmental estimates

All assigned confidence values for quality and precision of environmental estimates were standardised to maximum of 1 and combined (averaged) for each environmental variable (where only on estimate confidence was available as either quality or precision, that value was used instead).

Each environmental variable was weighted according to its importance as a predictor in the specific model for the species life stage, and the confidence in the combination of environmental data used by each model (C_e) was calculated in each grid cell as the weighted average following Equation 4.

Total confidence (spatial)

The total confidence variability within the map was calculated as before, by weighting the model performance score (C_m) by the mean confidence score of the input data (C_e only in this case), i.e.:

$$C(spatial) = C_m * C_e \qquad \text{(Eq. 8).}$$

This was represented as a relative confidence (higher to lower) to be read in relation to the overall confidence associated with the map as a whole.

A3.2 References

Czakan J. (2021) F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose? https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

Kay S., McEwan R. and Ford D. (2021) North West European Shelf Production Centre NWSHELF_MULTIYEAR_BIO_004_011. Quality Information Document (QUID), Copernicus Marine Environment Monitoring Service (CMEMS). Document ref. CMEMS-NWS-QUID-004-011, Issue 5.1, 24 February 2021, 49 pp. https://catalogue.marine.copernicus.eu/documents/QUID/CMEMS-NWS-QUID-004-011.pdf

Kubat, M., Holte, R.C. and Matwin S. (1998) Machine learning for detection of oil spills in satellite radar images. Machine Learning 30: 195-215. https://link.springer.com/content/pdf/10.1023/A:1007452223027.pdf

Lewis, D. and Gale, W. (1994) A Sequential Algorithm for Training Text Classifiers. Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 3–12), Springer-Verlag