ICES Journal of Marine Science

ICES International Council for the Exploration of the Sea
CIEM Conseil International pour l'Exploration de la Mer

ICES Journal of Marine Science (2016), doi:10.1093/icesjms/fsw136

Spatial prediction of demersal fish diversity in the Baltic Sea: comparison of machine learning and regression-based techniques

Szymon Smoliński* and Krzysztof Radtke

Department of Fisheries Resources, National Marine Fisheries Research Institute, Kołłątaja 1, Gdynia 81-332, Poland

**Corresponding author: tel: +48 587 356 193; fax: +48 58 73-56-255; e-mail: ssmolinski@mir.gdynia.pl.*

Smoliński, S. and Radtke, K. Spatial prediction of demersal fish diversity in the Baltic Sea: comparison of machine learning and regression-based techniques. – ICES Journal of Marine Science, doi:10.1093/icesjms/fsw136.

Received 9 April 2016; revised 29 June 2016; accepted 1 July 2016.

Marine spatial planning (MSP) is considered a valuable tool in the ecosystem-based management of marine areas. Predictive modelling may be applied in the MSP framework to obtain spatially explicit information about biodiversity patterns. The growing number of statistical approaches used for this purpose implies the urgent need for comparisons between different predictive techniques. In this study, we evaluated the performance of selected machine learning and regression-based methods that were applied for modelling fish community indices. We hypothesized that habitat features can influence fish assemblage and investigated the effect of environmental gradients on demersal fish diversity (species richness and Shannon–Weaver Index). We used fish data from the Baltic International Trawl Surveys (2001–2014) and maps of six potential predictors: bottom salinity, depth, seabed slope, growth season bottom temperature, seabed sediments and annual mean bottom current velocity. We compared the performance of six alternative modelling approaches: generalized linear models, generalized additive models, multivariate adaptive regression splines, support vector machines, boosted regression trees and random forests. We applied repeated 10-fold cross-validation, using accuracy as the measure of model quality. Finally, we selected random forest as the best performing algorithm and implemented it for the spatial prediction of fish diversity from the Baltic Proper to the Kattegat. To obtain information on the data reliability and confidence of the developed models, which are essential for MSP, we estimated the uncertainty of predictions with standard deviation of predictions obtained from all the trees in the ensemble random forest method. We showed how state-of-the-art predictive techniques, based on easily available data and simple Geographic Information System tools, can be used to obtain reliable spatial information about fish diversity. Our comparative work highlighted the potential of machine learning method to reduce prediction error in modelling of demersal fish diversity in the framework of MSP.

Keywords: Baltic Sea, demersal fish diversity, ecological modelling, ecosystem-based management, machine learning, marine spatial planning, random forest.

Introduction

Marine spatial planning (MSP) is considered a valuable tool in ecosystem-based management of marine areas (Caldow *et al.*, 2015). Different biological, physical, and socioeconomic criteria can be combined in this management approach (Young *et al.*, 2007). In the pursuit of ecological goals, MSP has the potential to minimize the disadvantages of single-species management and to reduce problems arising from discrepancies between ecological and jurisdictional boundaries (Crowder and Norse, 2008). Despite the benefits of MSP, management authorities encounter considerable technical challenges due to limited and inconsistent information about the environment (Foley *et al.*, 2010). Spatially continuous environmental data, such as the full-coverage maps, are required for proper decision-making. During the last decade, much effort was devoted to collecting data from a wide range of scientific fields, and many organizations have been involved in common databases (Caldow *et al.*, 2015). These joint efforts have provided valuable possibilities for the large-scale mapping of

biological communities and habitats, which can be helpful for decision-makers at least during the initial phase of management plan development (Joy and Death, 2004).

In recent years an increasing amount of studies presenting an integrative approach to biodiversity conservation has been observed (Stuart-Smith *et al.*, 2013; D'Agata *et al.*, 2014). In this approach, different aspects of diversity within species assemblages can be quantified (taxonomic, phylogenetic and functional diversity). However, taxonomic type of diversity, expressed, e.g. by species richness, is the most commonly used component of diversity and may be considered as a surrogate for other components (Devictor *et al.*, 2010). The role of ecologists is to provide, using their best scientific knowledge, quantitative and spatially explicit information about biodiversity patterns (Pittman *et al.*, 2007). However, even in intensively studied areas, the spatial distribution of species cannot be fully monitored due to technical limitations and economic reasons (Leathwick *et al.*, 2006a). A reliable and cost-effective technique used to fill in these inevitable gaps in the biological information is predictive modelling (Guisan and Zimmermann, 2000; Elith *et al.*, 2006). The aim of this approach is firstly to relate by the model the biological survey data from sampled sites to environmental predictors and secondly to provide a map of investigated measure across whole region of interest (covering also unsampled areas) based on prediction of developed model (Ferrier and Guisan, 2006).

Predictive modelling has been widely used for mapping species richness, diversity, biomass or the abundance of different groups of organisms, in both terrestrial and aquatic ecosystems (e.g. Joy and Death, 2004; Bucas *et al.*, 2013; Lopatin *et al.*, 2016). The statistical approaches used for the purpose include more traditional regression-based techniques, such generalized linear models, generalized additive models, and multivariate adaptive regression splines, as well as novel machine learning (ML) algorithms, such as support vector machines, boosted regression trees and random forests. Previous experience has shown that ML techniques can be much more flexible than conventional parametric models due to their ability to handle non-linear relationships and complex interactions, which often occur in ecological data (Guisan and Zimmermann, 2000). Notwithstanding, a review of the literature showed a relatively low number of ML applications in ecology compared with other scientific fields (Olden *et al.*, 2008). Several studies have highlighted the urgent need for comparisons of the performance of different ML predictive techniques in ecology (Aertsen *et al.*, 2011; Olaya-Marín *et al.*, 2013).

The ecology of fish seems to be a promising but little exploited field for the application of ML. This group of organisms is considered an effective indicator of aquatic ecosystem quality due to its sensitivity to anthropogenic disturbances (HELCOM, 2012; Smoliński and Całkiewicz, 2015). Fish species richness and diversity are often used as a primary measures of ecological shifts and as a basis for planning protected areas (Knudby *et al.*, 2010b; Olaya-Marín *et al.*, 2013). Knowledge of the relationships between fish assemblages and environmental factors is important for effective conservation, and the application of novel statistical techniques can improve our understanding of these ecological processes (Olden *et al.*, 2008). A number of predictive analyses on freshwater fish communities, employing ML approaches, can be found in the literature (Olaya-Marín *et al.*, 2013). Most available predictive applications of ML in marine fish ecology refer mainly to coral reef-associated assemblages in shallow waters (Pittman *et al.*, 2007; Moore *et al.*, 2009; Knudby *et al.*, 2010a; Pittman and Brown, 2011). Such works on demersal fish in other ecosystem types are very scarce (Leathwick *et al.*, 2006a; Monk *et al.*, 2010; Compton *et al.*, 2012; Bergström *et al.*, 2013; Bucas *et al.*, 2013).

The aim of this study was to evaluate the performance of conventional regression-based techniques and machine learning used for the predictive modelling of demersal fish diversity. It was assumed that habitat-driven attributes can determine fish assemblage diversity (Knudby *et al.*, 2010a, b; Leathwick *et al.*, 2006b; Pittman *et al.*, 2007). Thus, based on openly available data, we investigated the relationship between demersal fish (species richness and Shannon–Weaver Index) and the environment of the Baltic Sea. The best-performing algorithm was implemented for the spatial prediction of fish diversity from the Baltic Proper to the Kattegat, accounting for the uncertainty of the estimation. We showed how state-of-the-art predictive techniques and simple Geographic Information System (GIS) tools may be applied to obtain reliable spatial information about demersal fish community.

Material and methods

Study area

The Baltic Sea, located in Northern Europe, is a semi-enclosed sea and one of the largest brackish-water basins in the world (415 200 km^2). The study area was located in the south of the Baltic Sea, covering a few distinct sub-basins south of latitude 59°N: Kattegat, Great Belt, The Sound, Kiel Bay, Bay of Mecklenburg, Arkona Basin, Bornholm Basin, Western Gotland Basin, Gdansk Basin, Eastern Gotland Basin, northern Baltic Proper (Figure 1) (HELCOM, 2015). The shallow Danish straits are the only connection with the Atlantic Ocean, where inflows of saline oceanic waters occur. Consequently, a high salinity gradient can be observed within the study area (from ∼5 in the northern Baltic Proper up to 27 in the western Danish straits). The Baltic Sea is considered an area of low biodiversity compared with the open oceans and most freshwater environments (HELCOM, 2009).

Fish data

We used data collected during the Baltic International Trawl Survey (BITS) programme in the 1st and 4th quarter of the years 2001–2014. To minimize the effect of variation in gear, we included in the analysis surveys conducted with two standard bottom trawls: TV3 930 meshes (TVL) and TV3 520 meshes (TVS) (ICES, 2011). Fish data were collected by different scientific institutions located around the Baltic and combined by the International Council for the Exploration of the Sea (ICES) in the Database of Trawl Surveys (DATRAS), which is freely available through the ICES website. In the analysis, we included only hauls where all species in the catch were recorded. We merged two datasets: the exchange data, containing detailed haul information, and the data on catch per unit effort (CPUE) per species and length group for each haul, comprising an indirect measure of the fish abundance of every species in different length classes.

The scope of investigation was the alpha-diversity (Foley *et al.*, 2010) of the demersal fish community, so we determined the allocation of all fish species to different ecological groups according to the information available on the FishBase.org website (Froese and Pauly, 2015) and excluded all pelagic fish. Herring (*Clupea harengus*), defined on FishBase.org as benthopelagic, was also excluded from the analysis because its Baltic population is

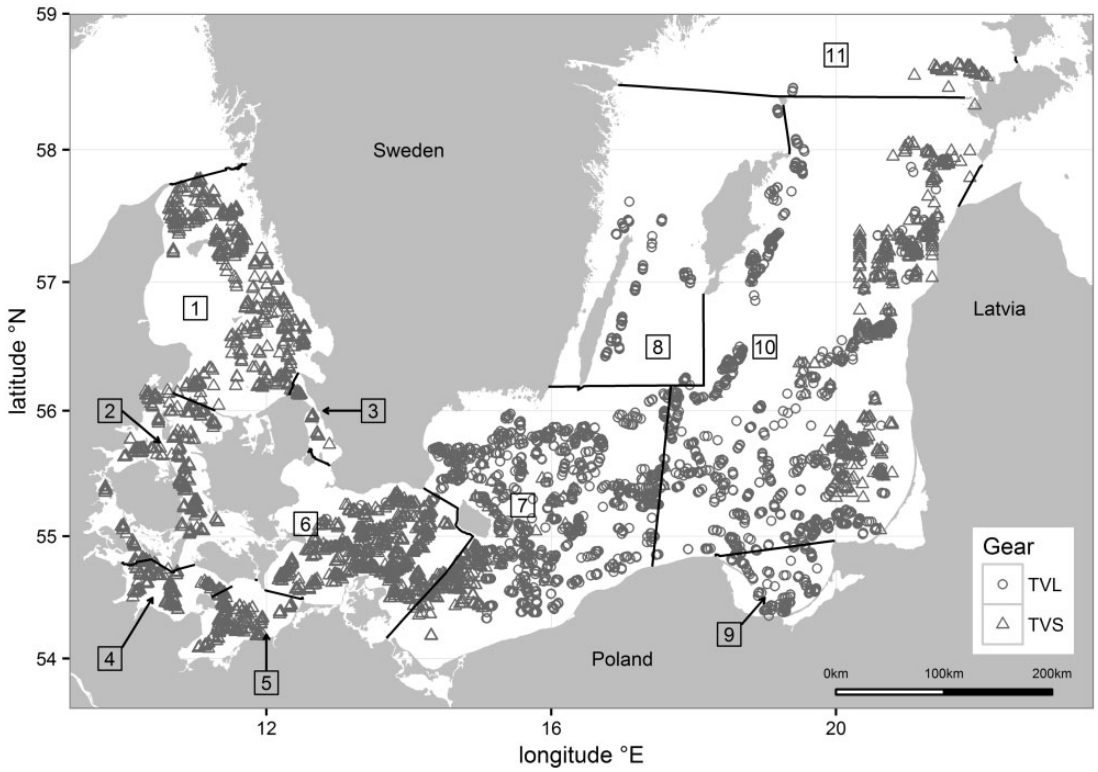

Figure 1. Fish sampling site locations. The shape of the points indicates the gear type used for trawling, lines show borders between HELCOM sub-basins: 1—Kattegat, 2—Great Belt, 3—The Sound, 4—Kiel Bay, 5—Bay of Mecklenburg, 6—Arkona Basin, 7—Bornholm Basin, 8—Western Gotland Basin, 9—Gdansk Basin, 10—Eastern Gotland Basin, and 11—northern Baltic Proper.

Table 1. List of predictors used for analysis of demersal fish diversity.

Predictors	Data type	Range	Data source
Haul data			
Quarter	Factor, 2 levels		(ICES, 2011)
Trawling gear type	Factor, 2 levels		(ICES, 2011)
Modelled environmental data (GIS layers)			
Bottom salinity	Factor, 5 levels		(Al-Hamdani and Reker, 2007; HELCOM, 2015)
Depth	Continuous	9.4–218.1 m	(Seifert *et al.*, 2001; HELCOM, 2015)
Seabed slope	Continuous	0–6.8%	(Al-Hamdani and Reker, 2007; HELCOM, 2015)
Growth season bottom temperature	Continuous	3.9–16.8 °C	(Al-Hamdani and Reker, 2007; HELCOM, 2015)
Seabed sediments	Factor, 5 levels		(Al-Hamdani and Reker, 2007; HELCOM, 2015)
Annual mean bottom current velocity	Continuous	0–0.13 m s^{-1}	(Al-Hamdani and Reker, 2007; HELCOM, 2015)

considered pelagic (Cardinale, 2000). In total, 88 species occurred in the dataset. The species diversity of the fish community was quantified using two indices: species richness and Shannon–Weaver (which combines species richness and evenness). We defined the species richness for each haul simply as the number of species in the catch. We calculated Shannon–Weaver Index of diversity based on the log $(x+1)$ transformation of the relative abundance of each species in the catch. Preliminary analysis showed that haul duration was a significant factor for observed species richness (increasing trend), so we included in the analysis only trawls with a standard duration of 30 min (5715 hauls in total) (Figure 1). Among the haul descriptors, we considered only the season of the survey (1st or 4th quarter) and the trawling gear type (TVL and TVS) as potential predictors of demersal fish diversity (Table 1).

Environmental data

We hypothesized that different features of marine habitats can influence the demersal fish community. We used maps of selected modelled physical features derived by the Baltic Interreg project BALANCE (Al-Hamdani and Reker, 2007) and a depth relief map (Seifert *et al.*, 2001) available from the HELCOM Baltic Sea data and map service (HELCOM, 2015) to provide environmental information for modelling demersal fish diversity and species richness. The selection of variables was based on the literature, expert knowledge and data availability. The list of predictors considered in the analysis is presented in Table 1. The GIS raster layers used in the study were resampled with bilinear interpolation or the nearest neighbour method to the cell size of 100 m and the same resolution was used for prediction. The range of GIS layers covered the whole Baltic Sea, but we intentionally reduced the

geographical scope of the analysis and set the northern edge to latitude 59°N, with regard to the distribution of fish sampling sites. Because no trawls were conducted in the regions of water bottom salinity specified originally in the model as Oligohaline I (salinity <5), we combined all areas with salinity <7.5 (originally classes Oligohaline I and Oligohaline II) into the first class (I) in our dataset. The maps used in the study are shown in the Supplementary Data for this article.

Data analysis

We performed all data exploration, development of different models, GIS analysis and predictive mapping using the *R* scientific computing language (R Development Core Team, 2011). We employed the Classification and Regression Training package *caret* (Kuhn, 2008) with the range of add-on packages for model building, tuning and accuracy assessment. Six alternative predictive techniques, listed in the following sections, were evaluated in this study. The *caret* package allowed us to train different algorithms in a consistent environment and to conduct direct comparisons of model performance. Moreover, internal tuning made it possible to optimize the model parameters, especially of the machine learning. We used 10-fold cross-validation resampling method to assess each model's accuracy (Hastie *et al.*, 2009). We additionally implemented a bootstrap approach by repetition of the whole cross-validation process 100 times with independent resampling of subsets, achieving a total of 1000 permutations for a single model. The results from the folds in each repetition were combined to select the model with the best parameters and to compare different algorithms. As the measure of model performance, we used the root-mean-squared error (RMSE) of the prediction estimated for each repetition. Because we conducted parallel tests for all modelling techniques, using the same test and training sets, detecting pairwise differences in the performances of alternative predictive techniques was a paired-sample problem (Knudby *et al.*, 2010a, b). Therefore, we applied pairwise comparisons using paired *t*-tests with the *p* value adjusted by the method of Benjamini and Hochberg (1995) to control the false discovery rate (FDR), the expected proportion of false discoveries amongst the rejected hypotheses. We evaluated the performance of the finally selected model by calculating RMSE of predictions for the whole study area and for each sub-basin.

Regression-based techniques

Herein, both generalized linear models (GLMs) (McCullagh and Nelder, 1989) and generalized additive models (GAMs) (Hastie and Tibshirani, 1990) were used for the predictive modelling of demersal fish community indices, using an appropriate family of distributions. Species richness data were treated using the Poisson distribution, while the Shannon–Weaver Index, due to its normal distribution, was analysed using a model based on the Gaussian law. We fitted separate models for both indices and all predictors, using, respectively, the basic *glm* function of *R* (R Development Core Team, 2011) and the *gam* function of the *mgcv* package (Wood, 2001). All variables were included in the analysis, since the generalized variance-inflation factors (GVIF) showed acceptable level (GVIF < 5) of predictors' collinearity. We used multiple smoothing parameter estimation by Generalized Cross Validation (GCV) with default gamma and 9 degrees of freedom allowed as maximum for the continuous predictors in GAMs. Additionally, Akaike's Information Criterion (AIC) was used by the *stepAIC* procedure of the *MASS* package (Venables and Ripley, 2002) for the stepwise feature selection of the GLMs.

We also applied multivariate adaptive regression splines (MARS) (Friedman, 1991) with implemented bagging procedure. We used the *bagEarth* function of *caret* based on the *earth* package (Milborrow, 2015). After preliminary tuning of the model, we set maximum number of terms in the pruned model to 14 and the maximum degree of interaction to 2.

Machine learning

In this study, support vector machines (SVM) (Cortes and Vapnik, 1995) based on the radial basis function (RBF) kernel was used. In the *caret* environment, the parameter σ was estimated using the *sigest* function of the *kernlab* package (Karatzoglou *et al.*, 2015), and the parameter C was tuned when running the algorithm.

To develop the boosted regression trees (BRT) (Elith *et al.*, 2008) we used the *gbm* package (Ridgeway, 2007). The values of the two BRT parameters were chosen from defined ranges in *caret* using a cross-validation method: the interaction depth (here 1, 3 or 9) and the number of trees (here 50–1500 in step of 50). We set the two remaining parameters manually as follows: minimum number of observations in the trees terminal nodes = 20 and shrinkage = 0.1.

The random forest (RF) (Breiman, 2001) model was developed using the *randomForest* package (Liaw and Wiener, 2002). We used cross-validation to check the performance of RF with one to eight variables randomly sampled as candidates at each split during the model building. The importance of variables was evaluated by the total decrease in node impurities from splitting on the variable, measured by the residual sum of squares and averaged over all trees.

Predictive mapping

Finally, we re-fit the best model with the whole dataset and optimal parameter values defined in the cross-validation process. We used this model further for predictive mapping of the demersal fish diversity indices, based on the GIS-layers of environmental data and fixed values of two haul factors (4th quarter and TVL gear type). In MSP activities, confidence in the predictions is crucial, and an appropriate presentation of uncertainty is important for decision-makers to understand data reliability (Caldow *et al.*, 2015). We obtained information about prediction uncertainty for each cell of produced map. In the case of linear models, bootstrap techniques can be used, while ensemble approaches such as boosted regression trees and random forests allow direct measurement of the standard deviation of predictions obtained from all the trees used for the final prediction (mean value of these trees). The coefficient of variation was calculated to provide a visualization of spatial regions with higher uncertainty levels. For that purpose, we used the *ModelMap* package, which makes it possible to read large GIS data in sections and maintain a reasonable usage of computer memory (Freeman *et al.*, 2010). This approach gave quantitative information on prediction validity, which was reduced into four qualitative categories (ranked according to the quartiles of the coefficient distribution) to enhance interpretability and map clarity.

Results

Fish assemblage indices

The species richness of the demersal fish communities in the analysed dataset ranged from 1 to 26 per haul (mean 5.8, standard deviation 4.2), and the Shannon–Weaver Index ranged from 0 to 3.03 (mean 1.39, standard deviation 0.66). We noticed a significant relationship between these two indices of the fish assemblages. For this reason, the comparison of all tested predictive methods gave the same model rankings for the Shannon–Weaver Index and species richness, with similar relative performance of particular techniques. Therefore, we decided to show only the results of the Shannon–Weaver Index modelling (see Species Richness Results in Supplementary Data).

Model performance evaluation

The predictive performance of the models is shown in Figure 2. The results of the repeated cross-validation indicate differences in the performance of the tested modelling approaches. ML methods, especially tree-based ensemble models, outperform regression-based techniques in terms of prediction accuracy. The differences in prediction errors obtained for each model by cross-validation were significant (the *t*-test adjusted *p* value for all models pairs was <0.001). The lowest RMSE was found for random forest, while the highest errors were observed for GLM. All tested models showed similar stability of the measured errors (the RMSE standard deviation ranged from 0.013 to 0.014).

Random forest

Random forest, as the algorithm with the best performance among the tested methods, was selected to predict the diversity of demersal fish. The results of the tuning process for the random forest parameter *mtry* showed that the optimal number of predictors randomly selected during the growing of each tree is 5.

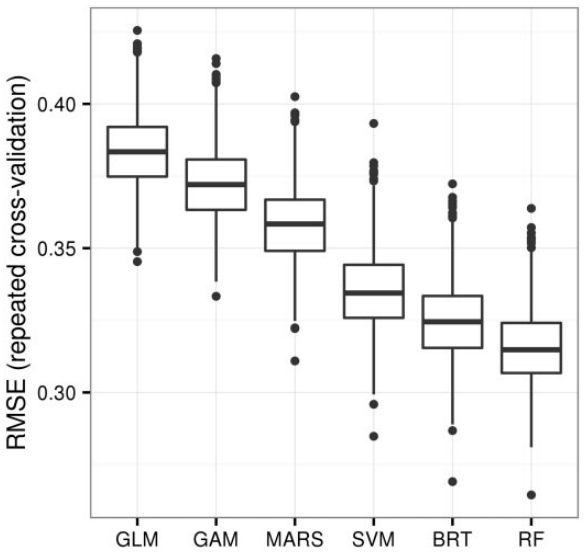

Figure 2. Root-mean-square error (RMSE) as predictive performance measure of the six selected modelling techniques for the Shannon–Weaver Index of the demersal fish community. Cross-validation (10-folds) with 100 repetitions was used for testing. Lines, boxes, and whiskers are medians, interquartile range (IQR), and 1.5× IQR of the estimated RMSE, respectively.

The addition of more variables does not increase the accuracy. The constructed random forest included 500 trees and explained 77.05% of the variance in the data. The model's predictive precision seemed to be better for the higher values of Shannon–Weaver Index, but slight underestimation was observed in the upper ranges. Conversely, the predicted values obtained from random forest were overestimated for observations when only one species was recorded and the Shannon–Weaver Index equalled zero. RMSE of predictions calculated for each sub-basin of the study area ranged from 0.08 in the Kattegat ($n=507$) to 0.31 in the northern Baltic Proper ($n=58$) (Table 2).

The salinity class was the most important predictor of demersal fish diversity (relative importance 42%), followed by depth (23%) and mean bottom temperature during the growth season (14%). The gear type and season of survey showed low levels of relative importance, 7% and 3%, respectively (Figure 3a). Partial dependence plots indicate that the Shannon–Weaver Index of the demersal fish community increases with salinity (Figure 3b). Furthermore, non-linear relationships can be observed between predictors expressed in a continuous scale (depth, bottom temperature, bottom current velocity, seabed slope) and response variables (Figure 3c–f). The local maximum of the Shannon–Weaver Index occurred in localities with ∼40 m depth and the highest observed bottom water temperatures. The partial dependence plots also show that fish diversity increases with increasing annual mean current velocity, up to ∼0.06 m s^{-1}, and remains stable for the higher values.

Predictive mapping

The modelled spatial distribution of the Shannon–Weaver Index highlighted a strong gradient of demersal fish diversity within the studied area, connected with high natural salinity differences in the Baltic Sea (Figure 4a). In geographic terms, the highest predicted Shannon–Weaver Index is observed in the Danish Straits and decreases with increasing distance from the Straits. The pattern of fish diversity distribution is also strongly influenced by the bathymetry. The lowest predicted Shannon–Weaver Index occurred in the deepest waters of the Baltic, and higher values are predicted along the coastline.

The prediction map is supported by the information on the uncertainty level obtained from the results of all 500 trees combined in the random forest model. The coefficient of variation was calculated for each cell in the map and ranked with

Table 2. Number of observations (*n*) and root mean square error (RMSE) of Shannon–Weaver random forest predictions calculated for sub-basins in the study area.

Sub-basin name	*n*	RMSE
Arkona Basin	975	0.10
Bay of Mecklenburg	202	0.16
Bornholm Basin	1541	0.16
Eastern Gotland Basin	1518	0.17
Gdansk Basin	140	0.17
Great Belt	358	0.13
Kattegat	507	0.08
Kiel Bay	169	0.16
Northern Baltic Proper	58	0.31
The Sound	72	0.11
Western Gotland Basin	175	0.18

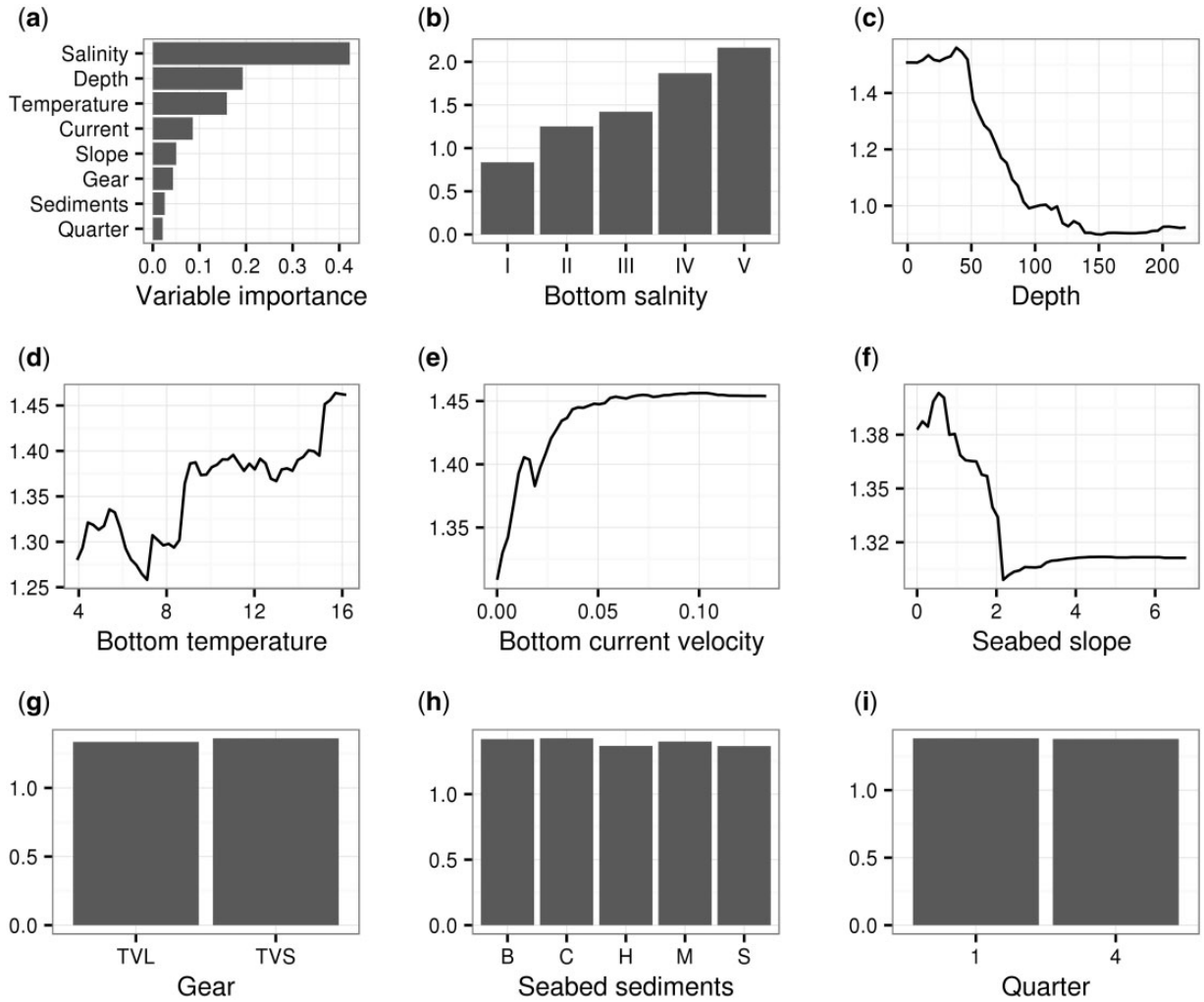

Figure 3. Relative importance of predictors expressed as % contribution in the model (a). Partial dependence plots for random forest regression of Shannon–Weaver Index and eight predictor variables (b–i). Partial dependence plots show the dependence of the response variable after marginalizing the effects of the other predictors. Variables are ordered by decreasing importance.

thresholds of classes calculated according to the quartiles of the coefficient distribution (Table 3). The final uncertainty map is presented in Figure 4b and shows that the prediction precision in the studied area varies due to unequal coverage of the collected samples and model abilities. The highest values of Shannon–Weaver Index were predicted with relatively low variation, which agrees with the results obtained during model evaluation. An unacceptable level of model precision (coefficient of variation >59%) occurred in the regions with low predicted diversity, especially the deeper areas, while in shallow waters much lower variance was observed.

Discussion

Predictive modelling has been widely used for mapping species richness in aquatic ecosystems by applying traditional regression-based techniques (GLM, GAM, MARS), but the application of machine learning in marine fish ecology is, to our knowledge, still surprisingly scarce. In particular, in the Baltic Sea area, comprehensive works on species modelling were conducted using maximal entropy (MAXENT), GAM and RF models to evaluate eutrophication management scenarios provided by the Baltic Sea Action Plan (Bergström *et al.*, 2013). The quoted studies, however, aimed to aggregate the forecasts of the applied methods rather than to compare them. The results showed that the combined models can be successfully used to predict perch (*Perca fluviatilis*) and pikeperch (*Sander lucioperca*) recruitment areas with respect to the changes in water transparency projected by eutrophication scenarios. The general findings of Bergström *et al.* (2013), indicating valuable properties of combined predictive techniques, outline further possible directions for the works presented herein. The ensemble approach can be used to minimize model-specific errors in predictions of species distribution.

Furthermore because the extent of predictive maps produced in our study was arbitrarily reduced with respect to sample coverage, more effort should be applied to testing the transferability of the developed models and their abilities to enlarge the mapped area. For example, the next step could be to test the models in areas further north in the Baltic Sea, outside the sampling coverage. The model applicability might differ considerably if the extension of the predicted area encompassed Baltic Sea bays, e.g. the Bothnian Bay, which constitutes a large part of the sea with a distinct environmental character (HELCOM, 2009). As shown by

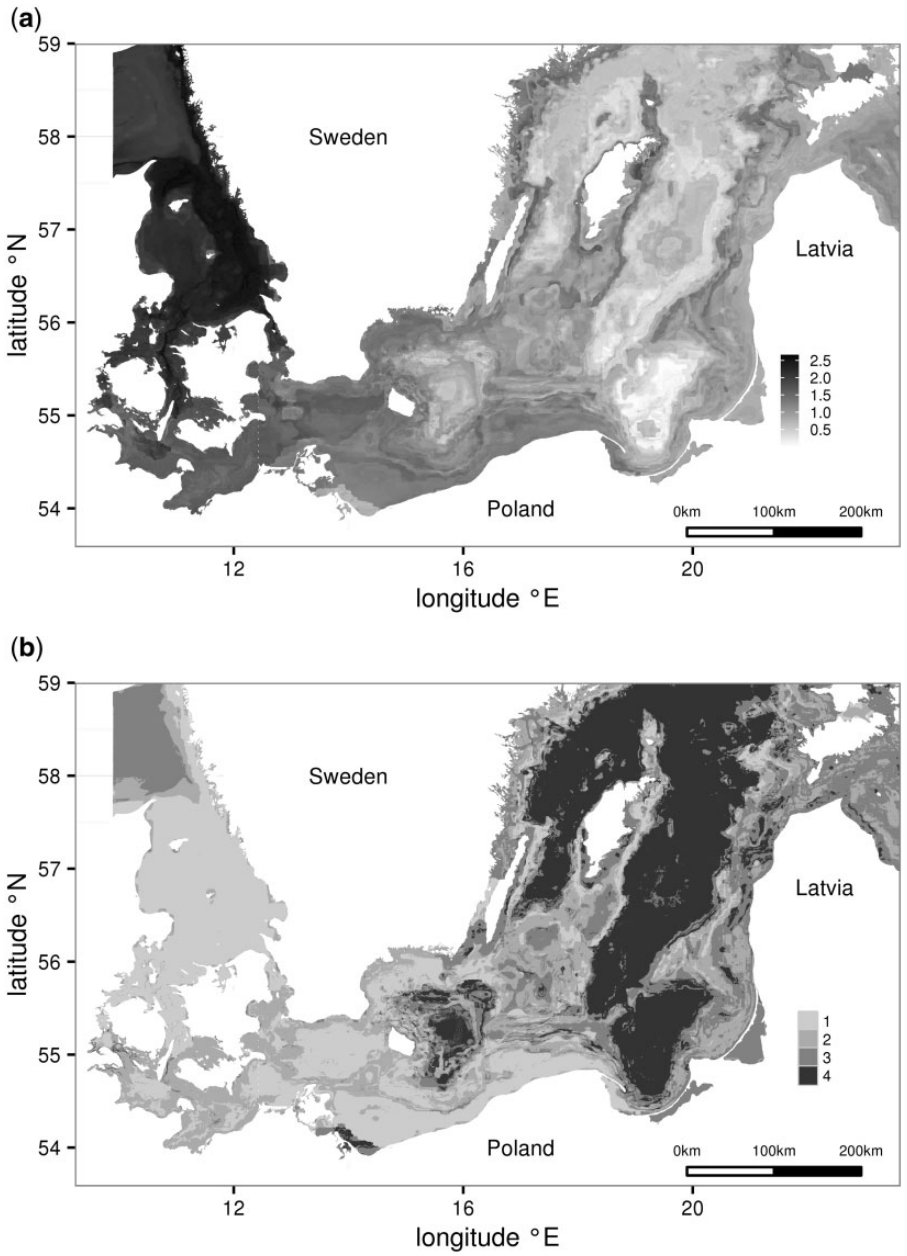

Figure 4. Map of predicted Shannon–Weaver Index of demersal fish community for the 4th quarter and TVL gear type using random forest model (a). Map of prediction uncertainty categorized with four levels, as specified in Table 3 (b).

Table 3 Thresholds for coefficient of variation used for assignment uncertainty levels of Shannon–Weaver Index prediction.

Coefficient of variation [%]	Rank of uncertainty
<28	1
28–41	2
42–59	3
>59	4

Sundblad *et al.* (2009), both differences in the observed range of the predictor variables in studied localities and the direct–indirect nature of the variables' effects should be considered in such investigations.

We identified the salinity class as the most powerful predictor of demersal fish diversity (relative importance 42%), followed by depth (19%) and mean bottom temperature during the growth season (16%). The remaining predictors were of low importance (<10%). Similar findings were presented by Leathwick *et al.* (2006a). They concluded that the high level of predictability was caused by the strong relationship between demersal fish species richness and the environment, with particular emphasis on depth and temperature. Depth of the oceanic region described in the paper by Leathwick *et al.* (2006a) varied from 5 to 1700 m. Remarkably diverse depth drives demersal fish species richness variability. Analogously, depth, coarse-scale topographic complexity measured by depth range or rugosity were the most important factors determining the diversity of coral reef-related fish

communities, although the depths did not exceed 8 m (Knudby *et al.*, 2010a, b). The results of the two quoted studies highlight depth as the strongest predictor, while in our study it was salinity, which results from the strong salinity gradient observed in the Baltic Sea, in contrast to the other discussed areas. It should be noted, that the depth effect in our study area may be masked by salinity, which varies considerably (approximately between 5 and 27) in comparison with stable salinity observed in the oceans surrounding New Zealand (differences <1) (Leathwick *et al.*, 2006a). Our results were in line with studies conducted by Snickars *et al.* (2014), who demonstrated high importance of salinity as predictor of fish distribution in the Baltic Sea.

Partial dependence analysis indicated a general increase in the Shannon–Weaver Index of the demersal fish community with increasing salinity, while for depth and bottom temperature, local maxima of the Shannon–Weaver Index were observed at the depth of ∼40 m and at the highest temperatures. Higher temperature and seasonal heat of water masses may lead to changes in stratification and the characteristic of food webs, e.g. by prompting biological production and increasing fish diversity (Mackenzie *et al.*, 2007). It is important especially in the case of the shallow waters of the Baltic and the Kattegat, warming up rapidly during the summer. The lowest diversity found in the deeps is caused by oxygen deficiency and anoxia, which affect fish distribution directly and indirectly by limiting prey availability (Mackenzie *et al.*, 2007). We may conclude that the environmental relationships of the Baltic Sea demersal fish diversity envisaged by the random forest method were mostly consistent with the available literature, showing high importance of hydrographical features as the predictors of fish distribution (Snickars *et al.*, 2014). However, Snickars *et al.* (2014) reviewed studies on species–environment relationships focusing on benthic organisms in the coastal areas of the Baltic Sea, finding hydrography (salinity and Secchi's depth) and biotic features to be the most widely used predictors of benthic fish. Their conclusions implied our suggestion to consider other potential environmental parameters in further model development: e.g. no information on water transparency or biotic traits was included in the predictive modelling presented herein. Water transparency in shallow waters is related to productivity and thus may affect both juveniles and adult fish by food supply (e.g. higher invertebrates density) (Snickars *et al.*, 2015). Besides the above-mentioned indirect effect, also direct impact is observed, as water transparency may affect the prey capture rate, predator avoidance or habitat choice of fish (Sundblad *et al.*, 2009).

From the methodological perspective, Lopatin *et al.* (2016) concluded that GLMs fitted for vascular plant species richness delivered better results in terms of precision and bias than RF, as GLMs can handle count data with better performance. However, in our case study, ML methods, especially tree-based ensemble models, surpassed regression-based techniques in terms of prediction accuracy. Similarly, results of studies by Bucas *et al.* (2013) pointed out RF as the most accurate method, when comparing with GAM and MAXENT. In general, the differences in prediction errors obtained for each model by cross-validation were statistically significant. The RF method revealed the lowest RMSE value among the models considered. In addition, a high level of agreement ($R^2 = 0.95$, $p < 0.001$) was demonstrated between the observed and predicted values of the Shannon–Weaver index for the RF model. The comparison of RMSE for RF models with different numbers of randomly selected variables (parameter *mtry*) with RMSE of other algorithms showed that the inappropriate tuning of model parameters can change the final decision on modelling method selection. For example, the use of $mtry = 3$ in RF building negatively influences the model performance and results in slightly worse accuracy than for the BRT method. The predictive precision of properly tuned RF seems to be better for higher values of the Shannon–Weaver Index. This feature of the fitted model is considered to be desirable, as the areas of higher biodiversity are of the greatest interest in a spatial management context (Roberts *et al.*, 2002). Differences in RF precision were also evident on sub-basins level. It was revealed by the lowest prediction error obtained for the Kattegat, where predicted fish diversity was the highest. In contrary, RMSE obtained for the northern Baltic Proper exceeded RMSE achieved for all the other sub-basins. It may result both from low number of observations ($n = 58$) and low diversity in the northern Baltic Proper. The robustness of the model and the statistical properties of the random forest indicated its selection for the prediction of demersal fish diversity in our investigation.

Any results of modelling should be interpreted with caution, taking into account the uneven distribution of sample coverage, which may have consequences in terms of prediction uncertainty (Young and Carr, 2015). The BITS are designed to obtain representative abundance indices of the most important bottom commercial species (cod and flounder). However, the data collected during the surveys give good opportunity to use them also for biodiversity modelling, as all species in the catch are recorded. The BITS haul allocation is random, but the coverage is constrained to trawlable areas (ICES, 2003). Consequently, some areas, e.g. with complex hard-bottom are permanently unsampled. Fish communities occurring on the areas of heterogonous seabed, which are underrepresented in the DATRAS, are commonly more diverse than soft-bottom assemblages. This implies that developed models may not be capable of handling variability of fish diversity in the small scale, governed mainly by local biotic factors (Florin *et al.*, 2009). Also deeps and shallow waters are not fully monitored (the distance between sampled locations may in some areas reach dozens of km). However, in the case of Baltic Sea deeps, it is assumed, that fish fauna is absent in oxygen depleted bottom zones (ICES, 2011). On the contrary, fish are usually more diverse in shallow waters, than in the open sea demersal zones (HELCOM, 2012), but the depths up to 10 m are not sampled during BITS, due to dimensions of the standard trawls and research vessels. In addition, the data used in the study are constrained to quarter 1 and 4. Consequently some fish migrations (e.g. spawning or feeding) are not incorporated in the presented model. The shortcomings of the data mentioned earlier, imply that time and spatial fish diversity patterns incorporated in the models are incomplete in the context of the whole year and in some Baltic Sea areas, but still may act as a useful surrogate for predicting fish diversity (Pittman *et al.*, 2007).

The environmental data used in the study were represented in the high spatial resolution, providing spatially detailed information about the distribution of selected physical features in the Baltic Sea (Al-Hamdani and Reker, 2007). Our approach was relatively simplistic, because we used static environmental variables that did not represent changes over several years, in which fish data were collected. For example, evident shifts in the distribution of anoxic and hypoxic waters have been observed during recent decades in the Baltic Sea, which might affect spatial patterns of fish diversity (Mackenzie *et al.*, 2007; Hansson and

Andersson, 2013). Moreover, some of the modelled environmental data do not take into account seasonal variability. The fish data from BITS survey were collecting during cold 1st and 4th quarter, while the information about modelled bottom temperature of water referred to the warm growth season. Furthermore, salinity as one of the most important factors affecting fish diversity was expressed in our study as discrete variable with five classes. The application of more detailed, continuous variable may improve accuracy of model. Besides the limitations, the use of these basic environmental data in the modelling of fish–habitat relationships allows to conduct prediction over broad spatial scale and obtain rapid and cost-effective information about the fauna distribution (Pittman *et al.*, 2007).

Our study showed the feasibility of applying novel statistical approaches to the spatial prediction of fish assemblage diversity. Based on the presented results, we share the opinion expressed in Bolker *et al.* (2013) that ecologists should be able to implement models both within and outside traditional statistical frameworks that seem suitable for their specific scientific investigations. This ability is particularly important now, as the dynamic development of natural resource management and conservation moves towards more spatial and ecosystem-based approaches (Crowder and Norse, 2008), which often require more flexible methods of biological data analysis (Elith *et al.*, 2008). The application of machine learning, such as random forests, may be valuable for obtaining knowledge on biotic and abiotic factors affecting species distribution and precise, quantitative information on spatial variation in species diversity, which are essential for management actions (Young and Carr, 2015). Our comparative work highlighted the potential of machine learning method to reduce prediction error in modelling of demersal fish diversity. More accurate models of marine fauna–environment relationships should improve predictive maps of species distribution used in MSP and in consequence provide more reliable qualitative basis for management of marine areas or strategies for biodiversity conservation.

Acknowledgements

This work resulted from the BONUS BIO-C3 project and was supported by BONUS (Art 185), funded jointly by the EU and the National Centre for Research and Development in Poland. We are grateful to U. Bergström and two other anonymous Reviewers for providing useful criticism on paper. We would like to thank also Piotr Margoński for valuable comments on earlier drafts of the manuscript and Piotr Potęga for the IT support.

Supplementary data

Supplementary material is available at the ICESJMS online version of the article.

References

Aertsen, W., Kint, V., Van Orshoven, J., and Muys, B. 2011. Evaluation of modelling techniques for forest site productivity prediction in contrasting ecoregions using stochastic multicriteria acceptability analysis (SMAA). Environmental Modelling and Software, 26: 929–937.

Al-Hamdani, Z., and Reker, J. 2007. Towards Marine Landscapes in the Baltic Sea. BALANCE interim report #10. Available at http://balance-eu.org/ (last accessed 24 May 2016). 118 pp.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B, 57: 289–300.

Bergström, U., Sundblad, G., Downie, A., Snickars, M., Boström, C., and Lindegarth, M. 2013. Evaluating eutrophication management scenarios in the Baltic Sea using species distribution modelling. Journal of Applied Ecology, 50: 680–690.

Bolker, B. M., Gardner, B., Maunder, M., Berg, C. W., Brooks, M., Comita, L., Crone, E., et al. 2013. Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS. Methods in Ecology and Evolution, 4: 501–512.

Breiman, L. 2001. Random forests. Machine Learning, 45: 5–32.

Bucas, M., Bergstrom, U., Downie, A. L., Sundblad, G., Gullstrom, M., von Numers, M., Siaulys, A., et al. 2013. Empirical modelling of benthic species distribution, abundance, and diversity in the Baltic Sea: evaluating the scope for predictive mapping using different modelling approaches. ICES Journal of Marine Science, 70: 1233–1243.

Caldow, C., Monaco, M. E., Pittman, S. J., Kendall, M. S., Goedeke, T. L., Menza, C., Kinlan, B. P., et al. 2015. Biogeographic assessments: a framework for information synthesis in marine spatial planning. Marine Policy, 51: 423–432.

Cardinale, M. 2000. Decreasing weight-at-age of Atlantic herring (Clupea harengus) from the Baltic Sea between 1986 and 1996: a statistical analysis. ICES Journal of Marine Science, 57: 882–893.

Compton, T., Morrison, M., Leathwick, J. R., and Carbines, G. 2012. Ontogenetic habitat associations of a demersal fish species, Pagrus auratus, identified using boosted regression trees. Marine Ecology Progress Series, 462: 219–230.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. Machine Learning, 20: 273–297.

Crowder, L., and Norse, E. 2008. Essential ecological insights for marine ecosystem-based management and marine spatial planning. Marine Policy, 32: 772–778.

D'Agata, S., Mouillot, D., Kulbicki, M., Andréfouët, S., Bellwood, D. R., Cinner, J. E., Cowman, P. F., et al. 2014. Human-mediated loss of phylogenetic and functional diversity in coral reef fishes. Current Biology, 24: 555–560.

Devictor, V., Mouillot, D., Meynard, C., Jiguet, F., Thuiller, W., and Mouquet, N. 2010. Spatial mismatch and congruence between taxonomic, phylogenetic and functional diversity: the need for integrative conservation strategies in a changing world. Ecology Letters, 13: 1030–1040.

Elith, J., Graham, C., Anderson, R., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R., et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29: 129–151.

Elith, J., Leathwick, J. R., and Hastie, T. 2008. A working guide to boosted regression trees. The Journal of Animal Ecology, 77: 802–813.

Ferrier, S., and Guisan, A. 2006. Spatial modelling of biodiversity at the community level. Journal of Applied Ecology, 43: 393–404.

Florin, A. B., Sundblad, G., and Bergström, U. 2009. Characterisation of juvenile flatfish habitats in the Baltic Sea. Estuarine, Coastal and Shelf Science, 82: 294–300.

Foley, M. M., Halpern, B. S., Micheli, F., Armsby, M. H., Caldwell, M. R., Crain, C. M., Prahler, E., et al. 2010. Guiding ecological principles for marine spatial planning. Marine Policy, 34: 955–966.

Freeman, E. A., Frescino, T. S., and Moisen, G. G. 2010. ModelMap : an R Package for Model Creation and Map Production. Available at: http://cran.r-project.org/web/packages/ModelMap/vignettes/VModelMap.pdf (last accessed 25 January 2016).

Friedman, J. H. 1991. Multivariate adaptive regression splines. The Annals of Statistics, 19: 1–67.

Froese, R., and Pauly, D. 2015. FishBase.org. www.fishabse.org (last accessed 3 December 2015).

Guisan, A., and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. Ecological Modelling, 135: 147–186.

Hansson, M., and Andersson, L. 2013. Oxygen Survey in the Baltic Sea 2013. Report Oceanography No. 49. Swedish Meteorological and Hydrological Institute, Göteborg, Sweden.

Hastie, T., and Tibshirani, R. J. 1990. Generalized additive models. *In* Monographs on Statistics and Applied Probability, p. 335. Chapman and Hall, London.

Hastie, T., Tibshirani, R. J., and Friedman, J. H. 2009. The Elements of Statistical Learning – Data Mining, Inference, and Prediction. Springer Verlag. 745 pp.

HELCOM. 2009. Biodiversity in the Baltic Sea. Helsinki: Helsinki Comission, 188 pp.

HELCOM 2012. Indicator-based assessment of coastal fish community status in the Baltic Sea 2005-2009. Baltic Sea Environment Proceedings, 131: 1–73.

HELCOM. 2015. Map and Data Service. http://maps.helcom.fi/website/mapservice/index.html (last accessed 11 April 2016).

ICES. 2003. Report of the Baltic International Fish Survey Working Group. ICES CM 2003/G:05 299 pp.

ICES. 2011. Report of the Baltic International Fish Survey Working Group, ICES CM 2011/SSGESST:05, Addendum 1: Manual for the Baltic International Trawl Surveys. 73 pp.

Joy, M. K., and Death, R. G. 2004. Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. Freshwater Biology, 49: 1036–1052.

Karatzoglou, A., Smola, A., and Hornik, K. 2015. Package 'kernlab'. Available at: https://cran.r-project.org/web/packages/kernlab/kernlab.pdf (last accessed 9 January 2016).

Knudby, A., Brenning, A., and LeDrew, E. 2010a. New approaches to modelling fish–habitat relationships. Ecological Modelling, 221: 503–511.

Knudby, A., LeDrew, E., and Brenning, A. 2010b. Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques. Remote Sensing of Environment, 114: 1230–1241.

Kuhn, M. 2008. Building predictive models in R using the caret package. Journal of Statistical Software, 28: 1–26.

Leathwick, J. R., Elith, J., Francis, M. P., Hastie, T., and Taylor, P. 2006a. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Marine Ecology-Progress Series, 321: 267–281.

Leathwick, J. R., Elith, J., and Hastie, T. 2006b. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling, 199: 188–196.

Liaw, a, and Wiener, M. 2002. Classification and regression by *randomForest*. R News, 2: 18–22.

Lopatin, J., Dolos, K., Hernández, H. J., Galleguillos, M., and Fassnacht, F. E. 2016. Comparing generalized linear models and random forest to model vascular plant species richness using LiDAR data in a natural forest in central Chile. Remote Sensing of Environment, 173: 200–210.

Mackenzie, B. R., Gislason, H., Möllmann, C., and Köster, F. W. 2007. Impact of 21st century climate change on the Baltic Sea fish community and fisheries. Global Change Biology, 13: 1348–1367.

McCullagh, P., and Nelder, J. A. 1989. Generalized linear models. *In* Monographs on Statistics and Applied Probability, p. 261. Chapman and Hall, London.

Milborrow, S. 2015. Earth: multivariate adaptive regression splines. *R Package*. Available at: https://cran.r-project.org/web/packages/earth/earth.pdf (last accessed 22 December 2015).

Monk, J., Ierodiaconou, D., Versace, V. L., Bellgrove, A., Harvey, E., Rattray, A., Laurenson, L., et al. 2010. Habitat suitability for marine fishes using presence-only modelling and multibeam sonar. Marine Ecology Progress Series, 420: 157–174.

Moore, C. H., Harvey, E. S., and Van Niel, K. P. 2009. Spatial prediction of demersal fish distributions: enhancing our understanding of species-environment relationships. ICES Journal of Marine Science, 66: 2068–2075.

Olaya-Marín, E. J., Martínez-Capel, F., and Vezza, P. 2013. A comparison of artificial neural networks and random forests to predict native fish species richness in Mediterranean rivers. Knowledge and Management of Aquatic Ecosystems, 409: 1–19.

Olden, J. D., Lawler, J. J., and Poff, N. L. 2008. Machine learning methods without tears: a primer for ecologists. The Quarterly Review of Biology, 83: 171–193.

Pittman, S. J., and Brown, K. A. 2011. Multi-scale approach for predicting fish species distributions across coral reef seascapes. PLoS ONE, 6: e20583.

Pittman, S. J., Christensen, J. D., Caldow, C., Menza, C., and Monaco, M. E. 2007. Predictive mapping of fish species richness across shallow-water seascapes in the Caribbean. Ecological Modelling, 204: 9–21.

R Development Core Team. 2011. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org.

Ridgeway, G. 2007. Generalized boosted models : a guide to the gbm package. Available at http://www.saedsayad.com/docs/gbm2.pdf (last accessed 10 January 2016).

Roberts, C. M., McClean, C. J., Veron, J. E. N., Hawkins, J. P., Allen, G. R., McAllister, D. E., Mittermeier, C. G., et al. 2002. Marine biodiversity hotspots and conservation priorities for tropical reefs. Science, 295: 1280–1284.

Seifert, T., Tauber, F., and Kayser, B. 2001. A High Resolution Spherical Grid Topography of the Baltic Sea, 2nd edn. Baltic Sea Science Congress, Stockholm, 25–29 November 2001, Poster #147.

Smoliński, S., and Całkiewicz, J. 2015. A fish-based index for assessing the ecological status of Polish transitional and coastal waters. Marine Pollution Bulletin, 101: 497–506.

Snickars, M., Gullström, M., and Mattila, J. 2014. Species-environment relationships and potential for distribution modelling in coastal waters. Journal of Sea Research, 85: 116–125.

Snickars, M., Weigel, B., and Bonsdorff, E. 2015. Impact of eutrophication and climate change on fish and zoobenthos in coastal waters of the Baltic Sea. Marine Biology, 162: 141–151.

Stuart-Smith, R. D., Bates, A. E., Lefcheck, J. S., Duffy, J. E., Baker, S. C., Thomson, R. J., Stuart-Smith, J. F., et al. 2013. Integrating abundance and functional traits reveals new global hotspots of fish diversity. Nature, 501: 539–542.

Sundblad, G., Harma, M., Lappalainen, A., Urho, L., and Bergström, U. 2009. Transferability of predictive fish distribution models in two coastal systems. Estuarine, Coastal and Shelf Science, 83: 90–96.

Venables, W. N., and Ripley, B. D. 2002. MASS: Modern Applied Statistics with S. Springer, New York.

Wood, S. N. 2001. mgcv: GAMs and generalized ridge regression for R. R News, 1: 20–25.

Young, M., and Carr, M. H. 2015. Application of species distribution models to explain and predict the distribution, abundance and assemblage structure of nearshore temperate reef fishes. Diversity and Distributions, 21: 1428–1440.

Young, O. R., Osherenko, G., Ekstrom, J., Crowder, L. B., Ogden, J., Wilson, J. A., Day, J. C., et al. 2007. Solving the crisis in ocean governance: place-based management of marine ecosystems. Environment, 49: 20–32.

Handling editor: Manuel Hidalgo