Incorporation of optimal environmental signals in the prediction of fish recruitment using random forest algorithms

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Canadian Journal of Fisheries and Aquatic Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>cjas-2017-0554.R1</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Article</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>19-Mar-2018</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Smoliński, Szymon; National Marine Fisheries Research Institute, Department of Fisheries Resources</td>
</tr>
<tr>
<td>Is the invited manuscript for consideration in a Special Issue? :</td>
<td>N/A</td>
</tr>
<tr>
<td>Keyword:</td>
<td>herring, sprat, climate, machine learning, data mining</td>
</tr>
</tbody>
</table>
Incorporation of optimal environmental signals in the prediction of fish recruitment using random forest algorithms

Szymon Smoliński

Department of Fisheries Resources, National Marine Fisheries Research Institute, Kollątaja 1, 81-332 Gdynia, Poland
tel.: +48 587-356-193, e-mail: ssmolinski@mir.gdynia.pl

Abstract

The drivers of recruitment of selected Baltic sprat (Sprattus sprattus) and herring (Clupea harengus) stocks were investigated. Data on the interaction dynamics among fish species, the biological characteristics of the stocks, the biomass of the main predators and the hydroclimatic environmental factors (Baltic Sea Index (BSI) and sea surface temperature (SST)) were used in the analysis. The combination of random forest (Boruta algorithm) and “sliding window” approaches was tested on the simulated data and then used for the selection of relevant predictors and the optimal time window for real environmental variables. SST had a significant positive effect on the recruitment processes. Moreover, contrasting effects were observed in the mean BSI from two different time windows. The same environmental variable generated contrasting short-term and long-term effects on fish recruitment. This paper highlights the potential benefits of random forest and data mining applications for the incorporation of environmental factors in the assessment of stocks. The proposed analytical approach may be valuable for the investigations of complex environmental impacts in a broad range of ecological studies.

Keywords

Baltic Sea, herring, sprat, climate, machine learning, data mining, sliding window
INTRODUCTION

Fish population dynamics depend strongly on recruitment processes (Cushing 1996), which were widely identified in 1980s as the crucial elements to be considered for effective management (Checkley et al. 2009). The recent availability of open-source environmental data provides an opportunity to test a variety of predictors that may have an influential role in stock recruitment. However, traditional statistical tests are often not flexible enough to analyze such complex data (Elith et al. 2008). Relationships between the number of recruits and environmental drivers are frequently nonlinear (Drinkwater 2005). Moreover, various interactions between drivers may occur in these ecological systems (Dreyfus-León and Chen 2007). To handle these challenges, machine learning (ML) techniques are used by researchers in many scientific fields (Olden et al. 2008). A variety of examples have previously shown that ML algorithms are more robust than traditional regression-based models (Guisan and Zimmermann 2000). ML algorithms have also been applied in predictive studies of fish stock recruitment (e.g., Bayesian networks in Fernandes et al. 2013, 2015; and artificial neural networks in Krekoukiotis et al. 2016). However, the applications of these algorithms in studies on fish ecology are still rare (Smoliński and Radtke 2017).

One of the ML techniques that has become increasingly popular is the random forest (RF) algorithm, which combines many single regression trees into an ensemble (Breiman 2001). Regression trees are used because they are efficient at selecting relevant predictors and can handle interactions (Elith et al. 2006). Moreover, regression trees can deal with different types of variables (e.g., categorical, numerical, dissimilarity matrices) and manage missing values with a minimal loss of information (De’ath 2007). A comparison of different models has shown the outstanding properties of RF in solving real world analytical problems (Fernández-Delgado et al. 2014).
The application of RF may be valuable for obtaining information on biotic and abiotic factors that affect marine fish (Smoliński and Radtke 2017). RF algorithm design makes it possible to estimate variable importance by simulating how much worse the prediction will be in the case of randomly permuted data for a selected predictor. Comparing the results for all predictors provides information on the relative importance of each variable (Prasad et al. 2006). A recent improvement of methods for variable selection is the Boruta algorithm (Kursa and Rudnicki 2010). This method was designed as an RF wrapper for identifying all relevant predictors, i.e., predictors shown by a statistical test to be more relevant than random probes. The Boruta approach appears to be useful for ecological investigations, in which the number of variables often exceeds the number of observations, thus hindering the robust selection of variables for a final predictive model.

If the variables used in the modeling are represented in a temporal scale, an additional selection of a time window for each predictor is necessary. A brief review of studies on the relationships between fish recruitment and climatic conditions revealed that a priori selected time windows are usually considered in the investigations. However, there are potentially many plausible competing environmental signal hypotheses (van de Pol et al. 2016). Recent studies have highlighted that climatic effects on organisms are likely complex and the identification of optimal signals may be a nontrivial task (Kruuk et al. 2015). A variety of potential time windows may be considered with “sliding window” analyses, which helps identify the optimal environmental signals via a statistically sound exploratory method (van de Pol et al. 2016). Because of the aforementioned advantages of RFs, their incorporation into the framework of “sliding window” technique appears to be promising step in the development of tools for the quantification of biological responses to environmental variability.
In this paper, the influence of different biological and environmental drivers on small pelagic fish recruitment will be considered using the example of main herring (Clupea harengus) and sprat (Sprattus sprattus) stocks in the Baltic Sea (ICES 2016). Environmental conditions have been suggested as determinants of fish recruitment via direct and indirect impacts; thus, effects may occur with different time lags. Random forests (Breiman 2001) and the Boruta algorithm (Kursa and Rudnicki 2010) were incorporated into the “sliding window” framework that was previously proposed for linear models (van de Pol et al. 2016) to test these phenomena in a systematic and statistical manner.

MATERIALS AND METHODS

Study area and species

The Baltic Sea is a semi-enclosed, shallow and brackish basin located in northern Europe (Fig. 1). The hydrological conditions within the Baltic are strongly dependent on sea water exchanges through the Danish Straits, which are the only connection with the Atlantic Ocean (Lehmann et al. 2002). A major reorganization of the Baltic ecosystem regime was observed in the late 1980s (Möllmann et al. 2009), and it was driven mainly by large-scale climatic forcing, which led to warmer and less saline conditions that considerably affected all trophic levels (Alheit et al. 2005). The trophic network of the Baltic Sea is relatively simple and presents three main fish species: cod (Gadus morhua), herring and sprat, which constitute ~80% of the total fish biomass (Margonski et al. 2010).

The subjects of the present study are two main Baltic pelagic stocks: i) herring in subdivisions of the International Council for the Exploration of the Sea (ICES) 25–29 and 32 (excluding 28.1, i.e., Gulf of Riga); and ii) sprat in subdivisions 22–32 (Fig. 1; ICES 2016). Previous studies have revealed that models of the relationships between spawning stock biomass and recruitment for Baltic herring and sprat explain only a small portion of the year-to-year
recruitment variability (e.g., Köster et al. 2003; MacKenzie and Köster 2004; Cardinale et al. 2009; Margonski et al. 2010). Thus, it was hypothesized that environmental drivers, such as hydroclimatic conditions and interactions with other fish species, as well as the biological characteristics of spawners may significantly affect the recruitment processes of these stocks.

Data

The recruitment abundance index (R; Fig. S1), the spawning stock biomass (SSB; Fig. S1) and weight at age groups (WAA; Fig. S2) of herring and sprat, and the total biomass (TSB; Fig. S3) of cod stocks in subdivisions 25-32 were obtained from the official ICES sources listed in Table 1. The R and SSB estimates for the period 1974–2016 were obtained from the results of virtual population analyses, which are calculated based on commercial catch data tuned by research-vessel survey data (ICES 2016). This method relies on assumptions of accurate reporting of catch data and information on biological parameters of the stock. Thus, the obtained estimates may be uncertain and could introduce variability to the analyses of stock recruitment (MacKenzie and Köster 2004). The weight by age groups was considered a proxy for fish condition, which for mature (spawning) individuals, is also significantly correlated with egg number (MacKenzie and Köster 2004). Information on the cod stock biomass was included in the analysis because of this species’ role as a main predator of pelagic fish in the Baltic Sea ecosystem (Cardinale et al. 2009).

Two additional environmental predictors represented at a higher temporal resolution were incorporated in the modeling of recruitment: sea surface temperature (SST; Fig. S4a) and Baltic Sea Index (BSI; Fig. S4b). Temperature data were obtained from the Extended Reconstructed Sea Surface Temperature (ERSST v4) model (Huang et al. 2015). The ERSST v4 is a monthly dataset produced on a $2^\circ \times 2^\circ$ grid, and it is spatially complete for global oceans. Average
monthly temperatures for the area 52-67°N, 10-31°E were used as a proxy of thermal conditions for the investigated stocks. In addition, the BSI, which is a proxy of water mass circulation within the central Baltic Sea, was used (Lehmann et al. 2002). This proxy is highly correlated with North Atlantic Oscillation but regionally adjusted to the characteristic of the Baltic Sea, and it is calculated as the normalized sea level pressure difference between Oslo, Norway (53°13′N, 14°13′E) and Szczecin, Poland (59°30′N, 10°30′E) (Fig. 1). Possible short-term direct influences (e.g., by physiological responses) and long-term indirect effects (e.g., by food supply) of both environmental variables were hypothesized; thus, different time lags were considered in the analysis.

**Data analysis**

The entire analytical process was conducted separately for herring and sprat using R as the response variable. For each stock, the set of explanatory variables covered the SSB and WAA of the modeled stock, the recruitment abundance and SSB of the competitive pelagic stock, and the TSB of cod (Table 1). Prior to the analysis, two missing cod TSB values were compensated for with an iterative imputation method based on a RF (Stekhoven and Bühlmann 2012). Because missing values were filled using the multicollinearity of the surrounding cells, this approach allows for the entire database to be used in the analysis and increases the statistical power without introducing spurious relationships between explanatory and response variables (Di Franco et al. 2016).

The general approach used in the present work for the identification of optimal environmental signals (Fig. 2) followed the “sliding window” technique proposed by van de Pol et al. (2016). The protocol of this stepwise systematic method allows for the rigorous identification of the best time window of predictors that are represented at a higher temporal
resolution. The first step of this analysis is the construction of a baseline model without the effects of the tested variables (e.g., hydroclimatic factors). Here, only biological variables were used to construct baseline model (Table 1). Then, a set of candidate models are listed by identifying all competing hypotheses (time windows of hydroclimatic factors) that require testing. Finally, comparisons between all models are conducted to identify the optimal predictor (variable with a critical time window) (Bailey and van de Pol 2016; van de Pol et al. 2016).

In the present study, absolute time windows were used to sequentially test each environmental predictor (SST and BSI), and the reference date was defined as the end of the spawning year. All possible combinations of time windows were considered within the range of 36 months before this date. For each time window, the arithmetic mean was used as an aggregate statistic of environmental conditions, and variables calculated in this manner were added to the set of biological predictors (Table 1).

Then, the Boruta algorithm (Kursa and Rudnicki 2010) and RFs were applied to the whole dataset in each time window to test the relevance of the added hydroclimatic predictors (BSI or SST). Original RF algorithm builds the regression trees based on bootstrapped samples, which are separated into training subsets (two-thirds of the samples), and remaining out-of-bag subset (OOB, one-third of the samples) (Breiman 2001). OOB is used for assessing the performance of the currently developed tree: to estimate the prediction error and to evaluate variable importance. Loss of model accuracy (the decrease in % of explained variance) from permuting the values in each variable is measured. Z-scores are then calculated by dividing the average loss of model accuracy by its standard deviation and used as importance measures for variables in the RF ensemble model. Boruta, designed as an RF wrapper, extends this idea and determines the relevance of variables via iterative comparisons of the importance of real
variables and random probes, which are called shadow attributes. They are created by shuffling the original variables in each algorithm iteration to form external reference for decision whether the importance of any given variable is statistically significant (discernible from importance which may arise from random fluctuation). For this reason, minimal, average and maximum Z-scores of added shadow attributes are calculated. Variables with significantly lower importance than maximum Z-score of shadow attributes are considered irrelevant (rejected), whereas those with higher importance are selected as relevant (confirmed) (Kursa and Rudnicki 2010). This extensive procedure was the basis for determining whether the environmental signal (mean SST or BSI from a specific time window) is the relevant predictor. Whole process have been repeated for each potential time window in the sliding window analysis. Because the number of Boruta iterations was limited to 500 for the sake of computational efficiency, certain variables in the model may be left without a decision (tentative). Such predictors were arbitrarily assumed to be relevant and included in the next step of the analysis. This decision had no effects on the results of environmental signal identification during sliding window stage.

Misclassification of environmental signal may occur in this step of the sliding window analysis by chance due to the large number of time windows considered. To evaluate directly the method performance, series of tests were conducted on the artificial datasets. The goal of the simulation was to assess the probability of false-positive and false-negative errors in the classification of environmental signal by Boruta algorithm. For this purpose 1000 artificial datasets, containing one response and one explanatory (environmental) variable, were generated for each level of known correlation between variables: $R^2 = 0$ (no environmental signal), 0.2, 0.4, 0.6, and 0.8 (strong environmental signal). Moreover, in order to consider differences in method performance under a variety of sample sizes, simulations with $N = 10, 20, 30, 42, 60, 100, 200,$
400, 800 and 1000 were carried out, giving 55000 simulation runs in total. In each run the Boruta
algorithm was used to classify environmental signal as true (relevant) or not. The rate of false-
negatives was calculated as the proportion of runs where dataset contained true signal ($R^2 \geq 0.2$),
but algorithm rejected environmental variable. The rate of false-positives was calculated as the
proportion of runs where the dataset contained no signal ($R^2 = 0$), but the algorithm classified the
environmental variable as relevant.

In the real example random forest models (Breiman 2001) were fitted for each time
window using relevant predictors identified by the Boruta method in the previous step. The
accuracy of the models was checked via 5-fold cross-validations repeated 100 times (Kuhn
2008), and the same training and test sets were used in each considered time window to make
results comparable. The root mean squared error (RMSE) metric was calculated as a predictive
performance measure of the models and used for optimal signal selection. The time window with
the lowest RMSE (the lowest unexplained variance of recruitment) was selected. Two iterations
of the entire “sliding window” analysis were conducted for each hydroclimatic variable, wherein
the second step models were refitted with incorporation of the first (optimal) determined signal.
Such an approach was motivated by the expected multiple signals of environmental variables,
which may drive fish recruitment with different time lags (e.g., short-term and long-term
effects), and it allows for the identification of secondary, suboptimal time windows of
environmental variables.

The four best-supported environmental signals for each stock (two time windows for each
of the two hydroclimatic predictors) were added to the dataset, and all predictors were retested
with the Boruta algorithm, assuming that possible interactions may occur between newly added
variables and cause changes in their relative importance. Then, the predictors selected by the
Boruta algorithm (confirmed and tentative) were incorporated into the final RF model. The predictive abilities of this model were compared with RFs based on biological variables only (Table 1), SSB and hydroclimatic signals or SSB only. Additionally, traditional Ricker stock-recruitment model and modified Ricker models with inclusion of identified hydroclimatic signals as controlling effects (Iles and Beverton 1998) were tested in terms of predictive performance. Curves were fitted using minimum least squares nonlinear regression. Optimal set of environmental factors in the modified Ricker models were selected based on the Akaike Information Criterion (AIC) obtained during the preliminary tests on the whole datasets. The 5-fold cross-validation was repeated 100 times, and the values of 25th 50th (median) and 75th percentile of the $R^2$ metric distribution were used for the comparisons.

RF algorithms generate additional information, which is often underutilized (Touw et al. 2013). For example, during the training of RF observations are recursively partitioned into the nodes of trees (Breiman 2001). Thus, the proximity between two observations can be calculated as the proportion of the time that the observations occur in the same terminal node of a tree to the number of trees in the entire RF ensemble (Cutler et al. 2012). This feature allows to obtain the proximity between samples and extract relevant trends from data with complex variable relationships (Touw et al. 2013). A principal component analysis (PCA) was conducted on the proximity matrix of the final model for all observations (recruitment years) in the dataset. Scores of the first two components were visualized to assess the similarities between recruitment years and for potential class discovery from RF clustering.

To evaluate the relationships between the response and selected environmental predictors, partial dependence plots were obtained, and they reflect the marginal effects of particular variables on the response within the RF model (Hastie et al. 2009). Prediction intervals were
calculated using the mean ± one and two standard deviations of all predictions obtained from individual trees in the RF ensemble (Freeman et al. 2010).

Throughout the analysis, RFs were grown with 500 trees. In the “sliding window” analysis, two-thirds of the variables available in each step were used for the construction of each regression tree, whereas in the final model, this parameter (mtry) was tuned via cross-validation. The Boruta (Kursa and Rudnicki 2010), randomForest (Liaw and Wiener 2002), ranger (Wright and Ziegler 2015) and caret (Kuhn 2008) packages in the R scientific computing environment (R Development Core Team 2011) were used in the study.

RESULTS

A slow decline of herring SSB was observed between 1974 and 2000 from approximately 1.7 to 0.4 million tons, and then an increase occurred in the remaining period up to 1 million tons in 2015. Herring recruitment at age 1 fluctuated at approximately 17 billion individuals during the investigated period and did not show a clear trend. The SSB of sprat increased between 1990 and 1996 to almost 2 million tons and then slowly decreased. Sprat recruitment at age 1 increased in the early 1990s similar to the SSB trend. Starting in 1995, a decreasing trend was observed, and then three strong recruitment years occurred: 2004, 2009 and 2015. The mean weight of individuals in all age groups decreased by approximately 30-40% between 1985 and 1995 for herring and by approximately 40-60% during the 1990s for sprat stock. The total biomass of the cod stock in ICES subdivisions 25-32 decreased during the 1980s from 1 million tons to approximately 0.16 million tons and fluctuated around this value until the end of the observed period.

According to the simulation results, the rate of false-positive classifications of environmental signal by Boruta algorithm was lower than 0.05 for all generated datasets (Fig. 3). The probability of such misclassification increases with increasing sample size up to N=30, but...
remain relatively stable for the datasets with more observations. The probability of false-negatives was above 0.05 for small datasets (10 samples) and correlations between $R^2 = 0.2$ and $R^2 = 0.6$. When simulated signal was low ($R^2 = 0.2$) in the datasets containing 30 or less samples, it was relatively often rejected by Boruta. In contrary, strong signal ($R^2 = 0.8$) was detected in most cases and even for the datasets containing 10 samples misclassification rate was on the level of 0.013.

The “sliding window” analysis conducted on the real data showed that the two best-supported time windows during which mean BSI influenced herring recruitment were as follows: i) 24-14 months before the reference date (end of the spawning year) and ii) the previous 16 months before this date (Table 2, Fig. 4a-b). The most significant effects of SST for herring stock recruitment were identified for the following periods: i) 9-5 months and ii) 22-21 months before the end of the spawning year (Table 2, Fig. 4c-d). The BSI at i) 11-9 months and ii) 19th months before the reference date and the SST at i) 4-3 months and ii) 32-23 months before the end of the spawning year were identified as the best environmental signals for predicting sprat recruitment (Table 2, Fig. 4e-h). The RMSE of the models in certain adjacent time windows was nearly as low as that of the identified optimal signals because of the temporal autocorrelation of hydroclimatic variables and similar sizes of their ecological effects (Fig. 4).

The results of the statistical tests conducted with the Boruta algorithm revealed that among the biological predictors and environmental signals created in the previous step, certain variables were relevant for the prediction of Baltic pelagic fish recruitment (Fig. 5). The most important predictors for herring were the optimal BSI signal, WAA 3, TBS of cod and recruitment of sprat. Boruta rejected the WAA 1, 8 and 5 from the herring model and the second signal of SST identified during the “sliding window” analysis (Fig. 5a). Similarly, the optimal
SST and both BSI signals were indicated as the best predictors for sprat recruitment success. High importance was also demonstrated for the sprat WAA 1-3 and 6. Boruta rejected the following variables: WAA 4 and 5, herring recruitment, herring SSB and second SST signal identified during the “sliding window” tests (Fig. 5b).

Tuning RFs (grown on a set of relevant variables identified by Boruta) with repeated 5-fold cross-validations showed that the optimal numbers of variables used to build each tree in the ensemble (mtry parameter) were 4 and 7 for the herring and sprat recruitment models, respectively. The PCA calculated on the proximity matrix of final RF models revealed major alterations in the recruitment processes between the years 1989 and 1990 for both investigated pelagic stocks (Fig. 6). The recruitment years before and after the shift formed two distinct groups on the PCA plots. One exception was sprat recruitment in 1976, which was more similar to the recruitment years after the identified reorganization (Fig. 6b).

The bivariate partial dependence plots derived for both BSI predictors used to construct the RFs revealed contrasting effects of different signals associated with the same variable. The long-lag effect of the optimal BSI signal (mean value from months 24-14 before the end of the spawning year) on herring recruitment was negative, whereas the short-lag effect of this variable (mean of the 16 months before the reference date) was positive (Fig. 7a). Similarly, the short-lag effect of BSI in the optimal window (11-9 months before the reference) had a positive effect on sprat recruitment, whereas the long-lag effect (19th months before the reference) was negative (Fig. 7c). In both cases, the threshold levels of BSI, which had a considerable influence on recruitment success, were observed. According to the model predictions, herring recruitment increases abruptly when the optimal BSI predictor drops below -0.1. A similar effect was observed when the mean BSI from the second time window exceeded 0.05. Sprat recruitment
increased sharply when the optimal BSI reached positive values and decreased when the mean BSI for the second time window was higher than -0.35. The recruitment of herring stocks showed a nonlinear positive response to the optimal SST signal, which was calculated as the mean value from the period March-July (Fig. 7b). For sprat, a positive relationship was observed between stock recruitment and the optimal SST (mean from the period August-September), with a clear threshold value observed at approximately 15.7 °C (Fig. 7d).

The final RF models for herring and sprat were significantly improved by the addition of environmental variables. The accuracy obtained in the cross-validations for models that considered all relevant variables (with identified hydroclimatic signals) exceeded the accuracy of the models that incorporated only biological variables and RF models of the relationship between recruitment and SSB (Table 3). However, RF model fitted for sprat on identified hydroclimatic signals and SSB revealed even higher predictive performance, than RF with all relevant variables (including also biological covariates). The observed improvement from the incorporation of hydroclimatic predictors was high for herring stock (from median $R^2=0.45$ to median $R^2=0.65$) and even higher for sprat (from median $R^2=0.10$ to median $R^2=0.51$). According to AIC values inclusion of two optimal hydroclimatic signals (first SST and first BSI) as controlling effects in the Ricker function was statistically supported for both stocks (Table S1, Table S2). Traditional stock-recruitment models obtained higher $R^2$ scores than RF if only SSB was taken into account as recruitment predictor. In contrary, RF models obtained higher predictive abilities than Ricker models when hydroclimatic variables were also included (Table 3).

The values predicted by the final RF models were underestimated in certain years when strong recruitment occurred (often referred to as spiked recruitment), especially in the case of the
sprat stocks (Fig. 8). All observed values ranged in the uncertainty area of ±1 standard deviations for all individual tree predictions in the ensemble model.

DISCUSSION

Relatively few ML applications can be found in ecology field compared with other disciplines (Olden et al. 2008). In particular, the relationships between recruitment success and environmental conditions have been investigated for many fish species using linear models (e.g., Cardinale et al. 2009; Margonski et al. 2010; Lindegren and Checkley 2013); however, applications of RF for these purposes are still rare (e.g., Hansen et al. 2015) despite the suitability of this modeling technique for such applied research because of its robustness and accuracy (Stekhoven and Bühlmann 2012). By incorporating bootstrap techniques, the RF algorithm shows good performance even when the number of predictors exceeds the number of observations (Díaz-Uriarte and Alvarez de Andrés 2006). The opportunity to test a large number of predictors within one model is considered a prominent feature of RF, and this algorithm may be especially helpful in recruitment modeling, where the number of observations (recruitment years) is often limited, thus preventing parameter estimations by more complex linear models with multiple predictors, such as Generalized Additive Models or Generalized Linear Mixed Models. Moreover, the general design of the RF algorithm helps obtain additional knowledge from the data, such as the variable importance for further selection of relevant predictors, proximity scores to investigate possible clusters of observations or partial dependence plots to visualize the main relationships between response and exploratory variables (Touw et al. 2013).

The advantages of the RF method have been used in this work for the selection of optimal hydroclimatic predictors of fish recruitment. The application of ML techniques for optimal signal identification was recommended by van de Pol et al. (2016), who developed a tool for “sliding
window” analyses within the framework of linear models. This exploratory method can be used to test different hypotheses (time windows) of environmental effects on ecological processes. In particular, variations in fish recruitment have often been considered an indirect response to changes in an additional driver (Drinkwater 2005). Thus, the a priori selection of only a single time window of potential predictors with limited knowledge, which is often the case in recruitment modeling, may miss relevant information. Although highly complex environmental influences on biological processes increase the difficulty of such analyses (Kruuk et al. 2015), the exploratory methods of signal identification presented here may facilitate the non-arbitrary selection of optimal time windows for investigated variables.

Misleading correlations between biological processes, such as recruitment and environmental data, are general statistical problem, since existence of a correlation does not imply the causation (Gulland 1952). Moreover, testing of large number of potential environmental signals (time windows) in the “sliding window” analysis leads to inevitable occurrence of false-positive signals, which can be found by chance. For this reason, the importance of conducting simulation tests to demonstrate, that novel method can correctly identify known signal is emphasized (Teller et al. 2016). Evaluation of method performance with known environmental signal and different sample sizes should be a first step towards identifying the potential pitfalls (Bailey and van de Pol 2016; van de Pol et al. 2016).

In this study datasets used for simulations were generated with assumption of simple linear relationships between two variables. No multiple interacting or confounding variables were taken into account (van de Pol et al. 2016). However, tree-based methods, like random forests, may handle such complex interactions between variables and non-linear relationships between them (Cutler et al. 2012). These features favor random forests as predictive technique,
giving extraordinary accuracy, but may also cause oversensitivity in detecting signals and patterns. Effective tests of signal relevance, that reduce rate of false-positive errors, is the general aim, which may be achieved with proposed application of Boruta algorithm. Boruta performs a top-down search for relevant variables. It compares original variable importance in the random forest model with importance of its randomized copies, providing unbiased and stable selection of important features (Kursa and Rudnicki 2010). Simulations conducted in the presented work revealed, that the probability of false-positive misclassification of environmental signal is relatively low, maintaining the level below 0.05. Also the rate of false-positives was on the low level for most of cases, showing high ability of signal detection by the random forests. However, because both types of errors seem to be unavoidable, results of environmental signal analysis based on proposed data-mining method should be taken with special care. Observational correlation studies are valuable tools to detect general ecological relationships and to generate new hypothesis, but more complete description of mechanisms of pelagic fish recruitment need further process-oriented experiments in the laboratory or in the field (Baumann et al. 2006).

R and SSB used in the presented study were obtained from virtual population analysis conducted based on commercial catch data, which are tuned by research-vessel survey data (ICES 2016). Reported landings and biological parameters assumed in the model can be imprecise and influence the stock assessment process (MacKenzie and Köster 2004). However, conducted analyses of residuals, retrospective patterns and sensitivity suggest appropriateness of the structural assumptions of Baltic herring and sprat assessment models (ICES 2016). Besides relatively high quality of representation of underlying populations and fishery processes, results of post hoc analysis based on R and SSB estimates derived by virtual population analysis, rather than raw data, should be taken with caution (Brooks and Deroba 2015). Cross-validation applied
in the presented study help to explore uncertainty associated with outlying data points with high leverage and assess predictive abilities of RF recruitment models giving a measure of the likely reliability of their application (Francis 2006). Additional tests repeated using confidence intervals of final estimates, results from stock assessment sensitivity analysis or retrospective runs are advised to address more directly different sources of uncertainty (Brooks and Deroba 2015).

By identifying multiple signals for the same environmental variable, different long-lag and short-lag effects on the investigated processes may be detected. The analysis revealed contrasting effects of the BSI from different time windows on the recruitment of herring and sprat. Positive BSI values in the period from August-December in the year preceding the spawning season or from January-March in the year of spawning had positive effects on the recruitment success of herring and sprat, respectively. High BSI values correspond to westerly winds over the Baltic, which transport warm and humid air masses to the Baltic region and result in higher sea surface temperatures (Lehmann et al. 2002; Möllmann et al. 2009), which may promote the reproduction of pelagic fish. Similar significant effects of relatively short-time signals of BSI on fish recruitment were also found in other studies of Baltic herring stocks (Cardinale et al. 2009; Gröger et al. 2014).

The results presented in this paper also indicate the possible occurrence of a long-term BSI signal, which has the highest importance in the herring recruitment model and was considered a suboptimal signal in the sprat recruitment model. Low BSI values had long-term positive effects on fish recruitment and the observed relationships varied with short-term BSI signals. Although interpreting such delayed effects identified by exploratory analyses may be difficult (Kruuk et al. 2015), these effects suggest the possible occurrence of indirect cascading
effects in the ecosystem through the succession of changes from climate to abiotic environment
to biotic environment to fish reproduction success, which may be delayed by long “signal
travelling times” (Gröger et al. 2014). Because low BSI and NAO values are correlated with
higher salinity in the Baltic Sea (Hänninen et al. 2000; Möllmann et al. 2003), it is assumed that
these saline and oxygen rich waters may stimulate the production of the ecosystem prior to the
spawning year and may influence spawner conditions, the further production of eggs, as well as
indirectly reduce mortality of the larvae and early juveniles.

The analyses performed here suggested that recruitment of the investigated stocks is
driven by temperature conditions. The SST from March to July in the spawning year was
considered one of the relevant predictors of herring recruitment success and had intermediate
importance. Depending on the locality, the spring spawning peak of herring in the Baltic Sea was
observed between March and July (Rajasilta 1993; Podolska et al. 2006), and the thermal
conditions during this period may have influenced the recruitment success of the stock. Cardinale
et al. (2009), Margonski et al. (2010) and Bartolino et al. (2014) have demonstrated that the
August SST has significant influence on herring recruitment. However, in these studies, only the
mean temperatures from the selected months were tested in the framework of Generalized
Additive Models. The number of hypothesized time windows of temperature variables were
limited because of the problems associated with candidate predictor collinearity or a priori
assumptions.

The SST from August to September in the year of spawning constituted the best predictor
of sprat recruitment. The temperature of the surface water masses during August was also found
to be the most significant predictor of sprat recruitment based on detailed correlation tests of the
stratified-by-depth monthly means of the thermal conditions in the Baltic Sea (Baumann et al.
2006). Voss et al. (2012) noted that sprat larval growth accelerates in May, is high in July and August, and slows from September onwards. These results may confirm the earlier conclusions that assumed a high contribution of individuals born late in the season in the Baltic sprat recruitment and the importance of processes observed during the late larval and early juvenile stages (Baumann et al. 2006). Results obtained in this study revealed, that temperature observed earlier in the year has also relatively high influence on sprat recruitment success. Peak spawning of sprat in the Baltic Sea is observed in April – June (Karasiova 2002) and is followed by 50-90 days of larval stage (Arrhenius and Hansson 1993). Thermal conditions in in the early part of the year may affect sprat recruitment e.g. by direct effects on sprat growth, mortality, maturation and egg production rates, as well as indirectly by reducing or stimulating spring production of zooplankton, which further constitutes essential prey for larvae and adult sprat (MacKenzie and Köster 2004; Baumann et al. 2006; MacKenzie et al. 2008).

The development of universal predictive models of fish recruitment for the entire investigated period (1974-2016) may be less efficient because of the regime shift of the Baltic Sea ecosystem that occurred in the late 1980s (Alheit et al. 2005; Möllmann et al. 2009). Abrupt alterations in the abiotic conditions of the Baltic Sea driven by phase transitions of the NAO and BSI caused the reorganization of the ecosystem structure and functioning (Möllmann et al. 2009), which promoted profound changes in the plankton communities and caused reactions of fish stocks by bottom-up mechanisms (Alheit et al. 2005). In the present study, the PCA conducted on the proximity matrix of the RF model revealed significant differences in herring and sprat recruitment in the years before and after 1989-1990. Reorganization of the ecosystem driven by large-scale climatic factors could have potentially changed the nature of the relationships between the recruitment of the studied species and environmental conditions.
Further modeling of stock recruitment should focus on the regime-specific drivers of these processes and possible differences in the relationships observed in each regime. Because RF algorithms are generally data-driven techniques, they may lose their predictive ability if values of variables go well beyond the range observed in the data (e.g. after strong alteration of environmental conditions). However, potentially RF may help to early identify changes in stock-recruitment-environment relationships by internal determination of similarity between samples (recruitment years) during development of regression trees ensembles (Touw et al. 2013). Preliminary analysis conducted on the presented datasets (not shown in this paper) suggested, that these properties of RF may be beneficial for identification of recent changes, but need to be tested in the future studies, taking into account different baselines from the longer time series, including various ecosystem states.

The structure of zooplankton communities and food supply affect the spawners and recruits of pelagic fish and are considered significant predictors of recruitment success (Köster et al. 2003; Cardinale et al. 2009; Raid et al. 2010). In this study, information on the feeding conditions were indirectly presented in the dataset as the mean weight of the individuals at age groups assuming a correlation between these variables. However, further inclusion of data on zooplankton availability for spawning individuals and early stages of fish may significantly improve the accuracy of the model. Moreover, only the two best signals of the variables were added sequentially based on the results of the optimal climatic signal detection. Because RF and many other ML techniques can handle correlated variables, robust predictions may be conducted based on a broader set of potential predictors, e.g., a high number of environmental signals coming from different time windows. A more restricted approach was applied in this study to
obtain a reasonable balance between the predictive accuracy and ecological interpretability of the model (Hansen et al. 2015).

One can expect that exclusion of tentative variables in the final step of the analysis may increase the accuracy of the prediction due to the elimination of noise (Kursa and Rudnicki 2010). Subsequent model building based on subsets of variables in the *minimal-optimal* fashion (Genuer et al. 2008) may also improve the predictive performance of RFs. Even though it was not the ultimate goal of this study, comparisons of the models showed, that RF of sprat recruitment limited to the hydroclimatic signals and SSB achieved better predictive results than the full model (including also biological factors). Nevertheless, developed RFs showed better performance than traditional and modified Ricker models. A main drawback of the majority of traditional methodologies for incorporating environmental factors into stock-recruitment models is the fact that investigated dependencies are usually non-linear and asymmetrical (Subbey et al. 2014).

This study demonstrated that the incorporation of environmental hydroclimatic variables into recruitment models significantly improved the prediction accuracy. More accurate recruitment predictions can advance annual stock assessments and advice; however, the number of fish stocks managed with respect to environmental conditions is still limited (MacKenzie et al. 2008). The current methods applied to perform recruitment estimates of the investigated Baltic pelagic stocks are based on a simple calibration regression and combination method (Shepherd 1997), and this approach neglects the variability of conditions in the marine environment (ICES 2016). Because data on environmental variables identified in the presented study as the best predictors of fish recruitment may be partly unavailable for forecasting during annual fishery advisory meeting due to the agreed schedule of works, future search of predictors may be
restricted to the period that give chance to obtain appropriate data in advance before assessment meetings and to implement them in the routine stock assessment procedures (MacKenzie et al. 2008). Such actions could increase the accuracy of predictive models of recruitment for year classes that will enter the fishery in the near future (Subbey et al. 2014), could help to provide such predictions earlier, giving management organizations and industry more time to adapt to changes in fish production (MacKenzie and Köster 2004) or could be applied in simulation studies to explore population dynamics under different scenarios of environmental change (MacKenzie et al. 2012).

In conclusion, the significant influence of SST on the recruitment processes of the investigated stocks was confirmed. Moreover, contrasting effects of the mean BSI from two different time windows were observed, which demonstrated that the same environmental variable may have different short-term and long-term effects on biological processes. These results highlighted that the incorporation of environmental factors in the assessment of stocks may improve fishery management procedures. Such actions can be especially beneficial in areas that are highly vulnerable to climate change, such as the Baltic Sea (HELCOM 2013), where the strong influence of climatic variations on fish population dynamics is clear (MacKenzie et al. 2007).

The combination of RFs, the Boruta algorithm and a “sliding window” approach was proposed for the selection of relevant predictors and identification of the optimal time window for environmental variables at a monthly resolution. The extension of available statistical techniques outside of traditional frameworks may improve the precision and accuracy of models and could be used to resolve more complex ecological questions (Bolker et al. 2013). This ability is particularly important now because the dynamic development of open-source environmental
and biological databases allow for testing a variety of hypotheses; however, such approaches require more flexible methods of statistical analysis. This paper highlights the potential benefits of data mining and ML applications in ecological modeling and shows that the proposed analytical approach may be valuable for the selection of relevant predictors and the investigation of complex environmental impacts in a broad range of ecological studies.

**ACKNOWLEDGEMENTS**

This work resulted from the BONUS BLUEWEBS project which has received funding from BONUS (Art 185), funded jointly by the EU, the Academy of Finland, Projektträger Jülich (PtJ), Germany, the State Education Development Agency of Latvia, the National Centre for Research and Development, Poland, and the Swedish Research Council Formas. The author would like to thank the associate editor and reviewers for their valuable comments on the manuscript and Professor Jan Horbowy for advice on traditional stock-recruitment-environment models.

**REFERENCES**


Atlantic fish species. Ecol. Inform. 25: 35–42. Elsevier B.V.

doi:10.1016/j.ecoinf.2014.11.004.


Table 1. List of variables used in the modeling of Baltic herring and sprat recruitment.

<table>
<thead>
<tr>
<th>Variable abbreviations</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>her/sprR</td>
<td>Herring or sprat recruitment at age 1</td>
<td>(ICES 2016)</td>
</tr>
<tr>
<td>her/sprSSB</td>
<td>Herring or sprat spawning stock biomass</td>
<td>(ICES 2016)</td>
</tr>
<tr>
<td>her/sprWAAx</td>
<td>Herring or sprat weight at age x</td>
<td>(ICES 2016)</td>
</tr>
<tr>
<td>codTSB</td>
<td>Total stock biomass of cod in subdivisions 25-32</td>
<td>(ICES 2013)</td>
</tr>
<tr>
<td>SST</td>
<td>Mean sea surface temperature</td>
<td>(Huang et al. 2015)</td>
</tr>
<tr>
<td>BSI</td>
<td>Mean Baltic Sea Index</td>
<td>(Lehmann et al. 2002)</td>
</tr>
</tbody>
</table>
Table 2. Optimal time windows for the hydroclimatic signals revealed during two iterations of the RF “sliding window” analysis. Windows are defined as the month before the reference date (end of spawning year), and calendar months are given with the year lag in brackets.

<table>
<thead>
<tr>
<th></th>
<th>Window open</th>
<th>Window close</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Herring</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BSI – 1&lt;sup&gt;st&lt;/sup&gt; step</td>
<td>24 (December-2 y)</td>
<td>14 (October-1 y)</td>
</tr>
<tr>
<td>BSI – 2&lt;sup&gt;nd&lt;/sup&gt; step</td>
<td>16 (August-1 y)</td>
<td>0 (December)</td>
</tr>
<tr>
<td>SST – 1&lt;sup&gt;st&lt;/sup&gt; step</td>
<td>9 (March)</td>
<td>5 (July)</td>
</tr>
<tr>
<td>SST – 2&lt;sup&gt;nd&lt;/sup&gt; step</td>
<td>22 (February-1 y)</td>
<td>21 (March-1 y)</td>
</tr>
<tr>
<td><strong>Sprat</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BSI – 1&lt;sup&gt;st&lt;/sup&gt; step</td>
<td>11 (January)</td>
<td>9 (March)</td>
</tr>
<tr>
<td>BSI – 2&lt;sup&gt;nd&lt;/sup&gt; step</td>
<td>19 (May-1 y)</td>
<td>19 (May-1 y)</td>
</tr>
<tr>
<td>SST – 1&lt;sup&gt;st&lt;/sup&gt; step</td>
<td>4 (August)</td>
<td>3 (September)</td>
</tr>
<tr>
<td>SST – 2&lt;sup&gt;nd&lt;/sup&gt; step</td>
<td>32 (April-2 y)</td>
<td>23 (January-1 y)</td>
</tr>
</tbody>
</table>
Table 3. Results of the cross-validation of the random forest and Ricker recruitment models that incorporate different sets of predictors. Prediction accuracy (median, 25th and 75th percentile of $R^2$ metric distribution) of the models are given.

<table>
<thead>
<tr>
<th>Model: variables included</th>
<th>Herring</th>
<th></th>
<th>Sprat</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25th</td>
<td>median</td>
<td>75th</td>
<td>25th</td>
</tr>
<tr>
<td>Random Forest: all relevant variables</td>
<td>0.464</td>
<td>0.646</td>
<td>0.744</td>
<td>0.321</td>
</tr>
<tr>
<td>Random Forest: SSB + hydroclimatic variables</td>
<td>0.224</td>
<td>0.384</td>
<td>0.558</td>
<td>0.463</td>
</tr>
<tr>
<td>Random Forest: biological variables</td>
<td>0.259</td>
<td>0.451</td>
<td>0.634</td>
<td>0.028</td>
</tr>
<tr>
<td>Random Forest: SSB</td>
<td>0.015</td>
<td>0.061</td>
<td>0.167</td>
<td>0.015</td>
</tr>
<tr>
<td>Ricker: SSB + hydroclimatic variables$^a$</td>
<td>0.409</td>
<td>0.421</td>
<td>0.429</td>
<td>0.262</td>
</tr>
<tr>
<td>Ricker: SSB</td>
<td>0.157</td>
<td>0.161</td>
<td>0.164</td>
<td>0.085</td>
</tr>
</tbody>
</table>

$^a$ - based on preliminary AIC comparisons only first SST and first BSI signals were included in the modified Ricker models.
Fig. 1. Map of the Baltic Sea with the indicated subdivisions of the International Council for the Exploration of the Sea (ICES). Stocks investigated in this study are i) herring in subdivisions 25–29 and 32 (excluding 28.1, i.e., Gulf of Riga) and ii) sprat in subdivisions 22–32. The map was created based on the layer of ICES statistical areas (ICES 2017).
Fig. 2. Flow chart of the analysis conducted separately for herring and sprat data.

77x67mm (600 x 600 DPI)
Fig. 3. Probability of misclassification of environmental signal calculated based on the 1000 generated datasets containing response and explanatory (environmental) variables with known correlation between them (colors). Different sample sizes (x-axis) were considered in the simulation. Level of 0.05 was indicated with the solid line. Log-scale was used on the x and y-axis. Data points at the lower bound of the plot indicate the zero misclassification rate.
Fig. 4. Results of the random forest “sliding window” analysis for the environmental effects on herring (a-d) and sprat (e-h) recruitment. Outcomes of the first (a, e, c, g) and second (b, f, d, h) steps of the optimal signal identification process for the BSI (a, b, e, f) and SST (c, d, g, h) are shown on the plots. The root mean squared errors (RMSEs) of random forest models for each time window are visualized with a color gradient. Only the RMSEs of models in which hydroclimatic factors were relevant (according to Boruta tests) are colored. The best-supported time window is marked with an x.
Fig. 5. Results of the Boruta tests of the relevance of predictors in the random forest model of herring (a) and sprat (b) recruitment. The distribution of importance estimates of variables obtained during Boruta runs are shown with boxplots. The median of the highest importance of the shadow attribute (shadowMax), which constitutes the reference level for the decision, is marked with a dashed line. Distributions of minimal and mean Z-scores of shadow attributes (shadowMin and shadowMean, respectively) provide additional information about the stochasticity of the information system. Boruta decisions on tested predictors are indicated with colors of boxplots.
Fig. 6. Plot of the first two principal component scores derived by the PCA based on a proximity matrix of the recruitment years of herring (a) and sprat (b). Times of major alterations in recruitment are indicated with arrows.

119x180mm (300 x 300 DPI)
Fig. 7. Partial dependence plots for the hydroclimatic variables for random forest predictions of herring (a, b) and sprat (c, d) recruitment. The effects of different BSI signals (according to results of the 1st and 2nd steps of the "sliding window" analysis; see Table 2) are presented.
Fig. 8. Plot of the observed (points) recruitment and predicted recruitment by the random forest model (open triangles) of herring (a) and sprat (b). Prediction uncertainty (±1 or 2 standard deviations of all individual predictions in the ensemble) is represented with shaded areas.