Ecological Indicators 101 (2019) 117–125

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Evaluating complex relationships between ecological indicators and environmental factors in the Baltic Sea: A machine learning approach

Annukka Lehikoinen[a,b,*], Jens Olsson[a], Lena Bergström[a], Ulf Bergström[a], Andreas Bryhn[a], Ronny Fredriksson[a], Laura Uusitalo[c]

[a] Swedish University of Agricultural Sciences, Department of Aquatic Resources, Institute of Coastal Research Skolgatan 6, 74242 Öregrund, Sweden
[b] University of Helsinki, Ecosystems and Environment Research Programme, Kotka Maritime Research Centre, Keskuskatu 10, 48100 Kotka, Finland
[c] Finnish Environment Institute (SYKE), Programme for Environmental Information. Latokartanonkaari 11, 00790 Helsinki, Finland

ARTICLE INFO

Keywords:
Bayesian network classifiers
Tree-augmented Naive Bayes
Entropy Minimization Discretization
Coastal fish communities
Baltic Sea

ABSTRACT

The state of marine ecosystems is increasingly evaluated using indicators. The indicator assessment results need to be understood in the context of the whole ecosystem in order to understand the key factors determining the status of these environmental components. Data available from the system's different components are, however, often heterogeneous: they may represent different spatial and temporal scales, and different parameters can be measured with different accuracy. This makes it difficult to evaluate the relationship between these variables and status of the environment using indicators. We studied whether probabilistic, machine learning-based classifiers could provide for assessing the relationships between multiple environmental factors and ecological indicators. This paper demonstrates the use of Bayesian network classifiers (Tree-augmented Naive Bayes classifier, TAN as the specific case example), used together with structural learning from data and Entropy Minimization Discretization (IEMD) algorithm to study environment-indicator relationships within coastal fish communities in the Baltic Sea. By using two Baltic-wide indicators of coastal fish community status and a heterogeneous set of potentially influential natural and anthropogenic variables, we explore and discuss the potential of the approach. Given pre-defined cutting points for the indicators, such as the classification thresholds of the indicator, the method enables identifying relevant variables and estimating their relative importance. This information could be used in environmental management to demonstrate at which threshold value the state of an indicator is likely to respond to a pressure or a combination of pressures. In contrast to many other multivariate statistical methodologies, the presented approach can handle missing data as well as data of varying types, from fully quantitative to presence-absence, in the same analysis.

1. Introduction

In order to restore and preserve the functioning of marine ecosystems and their biological diversity, we need to understand the effects of multiple natural and anthropogenic factors acting on the environment simultaneously. The state of marine waters is increasingly evaluated using indicators (Large et al., 2013; Kupschus et al., 2016; Teixeira et al., 2016; Tam et al., 2017), but the understanding of the relationships between these indicators and their associated environmental conditions, while deemed relevant (Rice and Rochet, 2005; Birk et al., 2012; Hattam et al., 2015), is often weak. In addition, the effects of numerous natural factors and anthropogenic pressures acting in concert are rarely addressed (Large et al., 2013; Uusitalo et al., 2016), despite the well-known prevalence of their interactive effects in biological systems (Large et al., 2013).

Studies evaluating the relationships between indicators and environmental factors often face analytical challenges related to addressing complex interactions among the factors in the environment. In addition, the possibility for analysis is typically constrained by the quality and precision of available data, as well as by shortages in data. Probabilistic machine learning approaches represent a potential solution, as they enable integration of data of different types and quality in an analytically coherent manner, and are able to deal with missing data points (Uusitalo, 2007; Barber, 2012).

In this study, we explore the potential of machine learning in the above described indicator-environment context. A set of potentially

* Corresponding author at: University of Helsinki, Ecosystems and Environment Research Programme, Kotka Maritime Research Centre, Keskuskatu 10, 48100 Kotka, Finland.
E-mail address: annukka.lehikoinen@helsinki.fi (A. Lehikoinen).

https://doi.org/10.1016/j.ecolind.2018.12.053
Received 23 August 2018; Received in revised form 18 December 2018; Accepted 31 December 2018
Available online 11 January 2019

influential natural and anthropogenic variables are tested as predictors of two ecological status indicators currently in use in the Baltic Sea area: *the abundance of coastal key fish species* (HELCOM, 2018a) and *the abundance of coastal fish key functional groups* (HELCOM, 2018b), included in Baltic Sea regional status assessment (HELCOM, 2018c). The first is represented by the abundance of European perch (*Perca fluviatilis*) in the central and northern Baltic Sea, whereas the second is composed of two parts; the abundance of piscivores and the abundance of cyprinids (*Cyprinidae*), the latter being in focus of our case study.

Having a central role in the coastal food web in the Baltic Sea (Östman et al., 2016), coastal fish communities (defined as the fish communities found in near-shore shallow waters, usually in waters shallower than 20 m) are currently included as one key element in national and regional assessments of the environmental and ecological status of the Baltic Sea, relating to the follow-up of the goals of the Baltic Sea Action Plan (BSAP, HELCOM 2007) as well as in the implementation of the EU Marine Strategy Framework Directive (MSFD, EC 2008). A wide range of environmental factors are known to impact these indicators, such as climate, trophic interactions and changes in water and habitat quality, which all vary on temporal and spatial scales (Bergström et al., 2016; Östman et al., 2017; HELCOM, 2018a,b). However, as for many other indicators used to assess the environmental status in the Baltic Sea, the contribution of multiple anthropogenic and natural drivers for the response of coastal fish indicators is not clear (HELCOM, 2018d). The issue is further complicated by the fact that data regarding these multiple factors are heterogeneous and often incompatible, some factors being quantitative and others semi-quantitative, categorical or qualitative.

We evaluated the performance of five different Bayesian network classifiers (Friedman et al., 1997) together with machine learning algorithms for finding optimal discretization and structure of the model to predict the state of the example indicators. The tested classifiers were the PC-algorithm (Spirtes et al., 2000), the Greedy search-and-score algorithm using both Akaike (Akaike, 1973) and the Bayesian information criterion (Schwarz, 1978), Naïve Bayes classifier (NB, Hand and Yu, 2001), and Tree-augmented Naïve Bayes (TAN, Friedman et al., 1997; Zheng and Webb, 2010). With the used data and evaluation criteria TAN outperformed the others, and its results are presented in this paper in detail.

This paper describes a protocol for using semi-supervised classification methods to increase understanding about the multiple pressures affecting an ecological indicator or some other response variable. We demonstrate the usefulness of the method in the environmental indicator context through analyzing simultaneously the contributions of both quantitative and semi-quantitative variables and finding meaningful threshold values for these variables without relying on linearity assumptions. At the end of the paper, we discuss the lessons learned, as well as the pros and cons of the approach.

2. Materials and methods

2.1. Data

The data originated from 41 fish monitoring areas in the northern Baltic Sea (Fig. A1; Table A1), 39 of them located along the east coast of Sweden and two in the Åland Islands and the Archipelago Sea areas of Finland. The areas cover substantial environmental gradients in salinity (2.2–7.5), nutrient concentrations (4–38 mg total phosphorus m^{-3} and 224–842 mg total nitrogen m^{-3}), and water temperature (summer temperatures 12–19 °C). They represent a mix of densely populated areas directly impacted by human activities, and reference areas with limited local anthropogenic impact (Bergström et al., 2016). The number of years monitored differed between the areas, from one to twelve years during the period 2002–2013 (Table 1, Table A1).

We used the area-specific average abundance of perch larger than 11 cm to represent the indicator *Abundance of coastal key fish species* (referred to as the *Perch indicator* hereafter), and the average abundance of cyprinids larger than 11 cm to represent the indicator *Abundance of coastal fish key functional groups* (referred to as the *Cyprinids indicator* hereafter) (HELCOM, 2018a,b). Data on 17 natural and anthropogenic variables in total, potentially affecting the abundance of fish were gathered from various sources, representing different aspects of water quality and hydrography, the availability of essential habitats, habitat degradation and natural mortality (Table 1). The ranges of the data are shown in Tables A2 and A3.

2.2. Analyses

Classification means predicting the class-level outcome of some variable of interest (in this case: whether an indicator attains a value above or below classification threshold for good environmental status) given information about the other variables (Friedman et al., 1997; Zheng and Webb, 2010). Bayesian network classifiers use the available data to build conditional probability distributions for the explanatory variables (usually and hereafter referred to as features; in our case study being the environmental variables) and the target variable that we want to predict (the class variable hereafter; i.e. the fish indicator variables in our case study). Using the Bayes rule, these probabilities can then be used to update the probability of the class variable given that the values of some or all of the features are known (e.g. Barber, 2012). All the analytical steps presented below were performed using the Bayesian network software Hugin (Educational 8.3) (Madsen et al., 2005).

2.2.1. Model framing and discretization of the variables

The indicator result is discrete, and most of the classification algorithms operate natively on discrete variables. Therefore, the data were discretized into multinomial values, each covering a distinct subrange of the original range of the continuous values. The discretization, however, simplifies the distribution and therefore necessarily causes loss of information (Uusitalo, 2007). In addition, the position of the class boundaries (i.e. *cut points*) affects the conditional distributions in the model and may have a strong effect on the results (Farnaz et al., 2017). To evaluate the influence of the boundary definition on the present models, multiple data files were created by discretizing the class variable in different ways. For each class variable (*Perch* and *Cyprinids indicators*), four data files were developed in which the data were divided equally into two to five classes (i.e. *bins*) (Fig. 1; these approaches are hereafter described as ED2-5). In addition, data files to study the conditions potentially leading to extremely high or extremely low indicator values were learned for the *Perch indicator* (Fig. 1; called ExtLow and ExtHigh hereafter). These cut points were defined by the domain experts. For the *Cyprinids indicator*, only the ExtHigh discretization was applicable due to the relatively high frequency of close to zero observations in the data (thus, an extreme low level could not be identified; see Fig. 1).

After the discretization of class variables following the above scheme, the features were discretized to minimize the information loss using the Information Entropy Minimization Discretization (IEMD) algorithm (Fayyad and Irani, 1993). IEMD discretizes the features so that their entropy given the discretization of the class variable is minimized, hence maximizing the predictive power of each feature. The cut point of the discretization learned by the IEMD processor thus indicates a point where some statistically meaningful change in the co-variation of the feature and class variables occur, given the discretization of the class variable. A feature was excluded from the model variant if the IEMD processor did not find any cut points for the discretization, as this implied that the variable was non-informative for predicting the state of the class variable.

2.2.2. Model selection

For all the alternative discretization approaches, we evaluated the performance of four different structural learning algorithms for

Table 1
Description of the data used in the analyses. Class variables (the fish indicators) are identified with an asterisk. HabPe was used for perch only and HabAv for the cyprinids only. Data range of each variable are shown in Tables A2 and A3.

Variable name	Explanation	Unit of measurement	Temporal coverage	Reference
Perch abundance* (Perch)	Annual estimate of abundance of perch (*Perca fluviatilis*) based on coastal fish monitoring using multi-mesh gillnets. The estimate represents average across all monitored stations within a sampling area.	Catch per unit effort of perch individuals ≥12 cm	Data availability varies between areas and years	Bergström et al. (2016)
Cyprinids abundance* (Cypr)	Annual estimate of abundance of cyprinid fish (mainly roach *Rutilus rutilus*, bleak *Alburnus alburnus*, bream *Abramis* spp., and rudd *Scardinius erythrophthalmus*) based on coastal fish monitoring using multi-mesh gillnets. The estimate represents average across all monitored stations within a sampling area	Catch per unit effort of all cyprinids individuals ≥12 cm	Data availability varies between areas and years	Bergström et al. (2016)
Salinity (Sal)	Salinity in the surface water (0–10 m) during Jan-Dec, averaged over the four most recent years	psu	Full availability for 2002–2013	SMHI (2016)
Total phosphorus (Ptot)	Total phosphorous concentration in the surface water (0–10 m) during Jun-Aug, averaged over the four most recent years	mg/m^3	Full availability for 2002–2013	SMHI (2016)
Total nitrogen (Ntot)	Total nitrogen concentration in the surface water (0–10 m) during Jun-Aug, averaged over the four most recent years	mg/m^3	Full availability for 2002–2013	SMHI (2016)
Temperature (Temp)	Summertime temperature in the surface water (0–10 m) during Jun – Aug, averaged over the four most recent years	°C	Full availability for 2002–2013	SMHI (2016)
Chlorophyll-a (Chl-a)	Chlorophyll-a concentration in the surface water (0–10 m) during Jun-Aug, averaged over the four most recent years	mg/m^3	Full availability for 2002–2013	SMHI (2016)
Sea water exchange (SWE)	Share of the open sea water in the surface layer (0–10 m) during Jun – Aug, averaged over the four most recent years	%	Full availability for 2002–2013	SMHI (2016)
Sampling temperature (TempF)	Annual estimate of the near-bottom temperature at each fish monitoring station. The estimate represents an average across all monitored stations within a sampling area	°C	Same years as fish monitoring	Bergström et al. (2016)
Sampling water transparency (TrF)	Annual estimate of the water transparency during fish monitoring. The estimate represents an average across all measurements within a sampling area.	m	Same years as fish monitoring	Bergström et al. (2016)
Mean depth (MD)	Mean water depth in the monitoring area, based on nautical chart data	m	Static over time	
Wave exposure (WExp)	Mean wave exposure in the monitoring area, obtained from a high-resolution wave exposure model	[index]	Static over time	Isæus 2004; Sundblad et al. (2014)
Distance to open sea (SDist)	Shortest water-way distance from the baseline (the inner limit for the territorial waters of a country) to the central point of each monitoring area, based on a cost distance analysis in ArcGIS	m	Static over time	
Latitude (Lat)	Geographical latitude of the central point of the monitoring area	Swedish national coordinate system RT90 2.5 gon V	Static over time	www.slu.se/kul
Perch habitats (HabPe)	Perch nursery habitat availability, mapped using species distribution modelling, in and nearby the monitoring area. Models based on extensive field data (n = 3969) on juvenile fish occurrence from 2005 to 2014	% of the modelled area	Static over time	Sundblad et al. (2013)
Coastal fish habitats (HabAv)	Averaged nursery habitat availability for roach, perch and pike (*Esox lucius*), mapped using species distribution modelling, in and nearby the monitoring area. Models based on extensive field data (n = 3969) on juvenile fish occurrence from 2005 to 2014	% of the modelled area	Static over time	Sundblad et al. (2013)
Cormorants (Corm)	Predation by cormorants (*Phalacrocorax carbo sinensis*), estimated from nest count data combined with consumption data, using a kernel density function in ArcGIS to interpolate consumption rates over the areas within feeding flight distance from the colonies	kg/km^2/season	Estimates available for 2006–2012	Finnish and Swedish national monitoring data on nesting great cormorants in 2006 and 2012 (linear interpolation used for 2007–2011)
Sticklebacks (HabSt)	Three-spined stickleback (*Gasterosteus aculeatus*) nursery habitat availability, mapped using species distribution modelling, in and nearby the monitoring area. Models based on extensive field data (n = 3969) on juvenile fish occurrence from 2005 to 2014. The variable is a proxy for predation and food resource competition with juveniles of the target species)	% of the modelled area	Static over time	Bergström et al. (2015), Sundblad et al. (2013)
Jetties (Jet)	Density of jetties in the monitoring area counted from orthorectified aerial photos	n/km^2	Static over time	Törnqvist and Engdahl (2010), Sundblad and Bergström (2014)

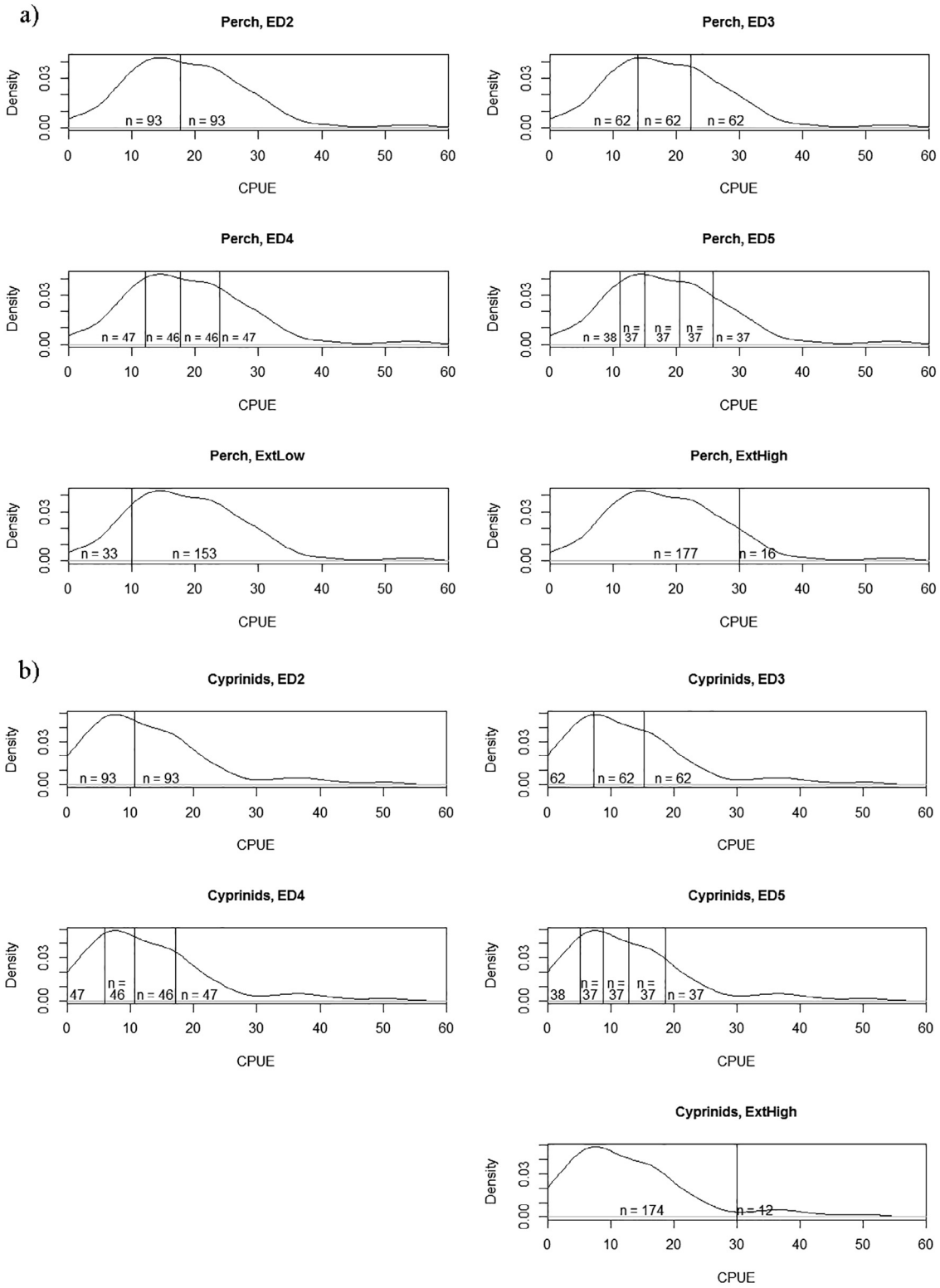

Fig. 1. Distribution of the (a) perch and (b) cyprinid indicator data used in this study (catch-per-unit-effort (CPUE) used as the measure of fish abundance) and the discrete class boundaries for the tested discretization cases.

Bayesian network classifiers in finding a model structure that 1) predicts the indicator result most accurately and 2) somewhat logically represents the known interrelations among the variables in the studied coastal ecosystems. The first criterion was evaluated using a set of performance metrics described below (Section 2.2.2), the results of the algorithm-wise comparisons being provided in the appendix (Tables A4 and A5). The evaluation of the second criterion was expert-driven, done by manual pairwise testing of the behavior of each variable when the state of the other variables were manipulated, checking that the direction of the well-known correlations are represented correctly by the model.

The tested algorithms were the constraint-based PC-algorithm

(original version developed by Spirtes et al. (2000)), the score-based Greedy search-and-score algorithm, using both Akaike (Akaike, 1973) and the Bayesian information criterion (Schwarz, 1978) and Tree-augmented Naïve Bayes (TAN, Friedman et al., 1997; Zheng and Webb, 2010). For comparison, the simplistic and fixed Naïve Bayes-structure was tested, too (Hand and Yu, 2001). The TAN algorithm was selected in this case for further analyses as it most often produced the best models in terms of the two criteria used. In the algorithm comparisons, the highest weight was given to the ED2 model of Perch indicator and ED3 model of Cyprinids indicator as they represent the real classification types of these indicators, the high abundance of perch (low – high-classification) and moderate abundance of cyprinids (low – moderate – high-classification) indicating the good ecological status.

The selected TAN classifier is an extension of Naïve Bayes (NB) classifier that has proven to perform well in various classification tasks (Hand and Yu, 2001; Zhang, 2004; Kuncheva, 2006; Ashari et al., 2013). While NB models assume that all of the features are dependent on the class variable only, and independent from each other, TAN allows a restricted amount of links between the features. These links are restricted so that (1) they must follow the basic Bayesian network principle and not form a directed loop, and (2) no more than one additional incoming link is allowed for each feature (Zheng and Webb, 2010). It has to be kept in mind that the aim of the machine-learned statistical classifiers is not to realistically present all the connections and correlations between the variables in the data, but to predict the value of the indicator as accurately as possible given the available information. Due to the limited amount of data available, complicated models are prone to overfitting the parameters to the available data. TAN model allows taking into account the strongest interactions between the explaining variables, while avoiding the risk of overfitting. In this study these links were derived from data using the Chow-Liu algorithm (Chow and Liu, 1968), the resulting probability distributions of the model being learned from data using the expectation maximization (EM) algorithm (Dempster et al., 1977; Lauritzen, 1995).

2.2.3. Model performance and functioning

The performance of Bayesian classifiers can be assessed in different ways (Korb and Nicholson, 2010). Usually, three different metrics are used for this purpose; two different error rates and additionally a multiclass extension of the Area Under Curve (AUC; e.g. Landgrebe and Duin, 2006). All of these metrics measure how well the model predicts the state of the class, but each metric addresses a specific aspect of the performance of the model. The error rate (ER) is the proportion of incorrect predictions in the model, where smaller values indicate a stronger prediction (James et al., 2013). In this study, error rates were calculated in two ways; by teaching the model with the full data and evaluating how many of the data points it would predict correctly (ER1), and by a leave-one-out cross-validation (ER2). The latter was run by teaching the model with all data but one and predicting this data point, repeated for every data point (n = 186).

AUC was used to summarize information contained in the Receiver Operation Characteristics (ROC) curve, which illustrates the performance of the model by plotting the true positive prediction rate against the false positive prediction rate. A perfect model has AUC value of 1, whereas a model with AUC 0.5 predicts correctly in 50% of cases, which corresponds to random guessing, thus meaning the model has no predictive power (Murphy, 2012; Flach, 2012; James et al., 2013). If the error rates of two models are similar but AUC values differ, the model with higher AUC performs better over all the operating points (Landgrebe and Duin, 2006).

Finally, the behavior patterns of the models were examined using an entropy reduction -based sensitivity analysis. The mutual information of two variables, i.e. to what extent information about one variable helps to predict the value of another, can be measured through the reduction of entropy in the distribution of the second variable when the first one is known. A discrete distribution has maximum entropy when its distribution is uniform, i.e. when each discrete category is equally likely (Murphy, 2012). An entropy reduction in the distribution of the class variable means that the uncertainty of the prediction is decreased. If the probability distribution of the class variable changes strongly when the value of the feature changes, the class variable is sensitive to this feature; for example, if the indicator value changes strongly when the water temperature is observed, the indicator is sensitive to water temperature and water temperature thus has high informative significance when the indicator is predicted.

One way to obtain accurate information about the functioning of the produced Bayesian network model and to study the dependencies in the data, is to use the model for different types of reasoning tasks (Korb and Nicholson, 2010). In practice the effects of updated information about some variable(s) on the probability distributions of the other variables are analyzed case specifically. This way the model can also be used for prediction tasks, if aimed to. In Section 3.4. a demonstrative example of the diagnostic use of one of the resulted TANs is provided.

3. Results

3.1. Model framing

The features (i.e. the explanatory environmental factors) included and excluded in the differently discretized data files and the corresponding model variants are presented in Table 2. The resulting discretization is presented for each specific model and feature in Tables A2 and A3. For the ED models, the number of features included tends to decrease as the number of bins in the class variable increased.

For the *Perch indicator*, the ED2 model consisted of nine features out of the 16 tested, while the ED5 model only included two (Table 2). A similar pattern of decrease occurred in the ED models of the *Cyprinids indicator*, but less pronounced. The numbers of features varied from 13 in the ED2 model to ten in the ED4 and the ED5 models (out of 16). Two features were included in all the ED models for the *Perch indicator*; predation by cormorants (*Corm*) and concentration of total phosphorus (*Ptot*) (Table 1; Table 2). For the *Cyprinids indicator*, eight of the features were included in all the ED models (Table 2).

The ExtLow model of the P*erch indicator* included only three features, two of which (salinity, *Sal*) and water transparency during monitoring, *TrF*) were not included in any other model (Table 1; Table 2). In the ExtHigh model variant of this indicator, no cut points were found for any of the features in the discretization phase, and the model was hence discarded. The ExtHigh model of the C*yprinids indicator* included six features, one of which (distance to open sea, *SDist*) was not included in any other model of the indicator.

In addition to the number of features included, some clear differences between the indicator models could be observed in the feature selection. For example, predation of cormorants (*Corm*) as found to be informative for discretization in all ED models of the *Perch indicator*, was not represented in any of the models of the *Cyprinids indicator*. Density of jetties (*Jet*) and sea water exchange (SWE) were included in almost all models of the *Cyprinids indicator*, but in none of the *Perch indicator* models.

3.2. Discretization of the features

Model-specific cut points of the different features are presented in Tables A2 and A3. Interesting feature-specific patterns could be identified. For some features, identical or very similar cut points were found repeatedly across models. For both the *Perch* and C*yprinids indicator* models, the cut point for mean depth of sampling area (*MD*) was consistently found around nine meters (Table A2; Table A3). The two cut points of total phosphorus concentration (*Ptot*) in the *Perch indicator* models were always found around 16 ± 1 mg/m^3 and 26 ± 1 mg/m^3. In three of the C*yprinids indicator* models, the cut point 5.3 mg/m^3 was identified. The cut points for salinity (*Sal*) were always found around

Table 2
Environmental variables (features) included in different discretization cases of the Perch and Cyprinids indicators (class variables) after feature discretization by the IEMD processor. The average informative significance of the feature for the indicator in TAN models is presented. The value presented is the reduction of the entropy (%) of the indicator's distribution, if only the feature in question is observed. The total number of features included in each model is shown on the last row (N feat.). For full names of the variables, see Table 1. "na" = "not applicable" (HabPe was used for perch only and HabAv for the cyprinids only.)

	Perch						Cyprinids				
	ED2	ED3	ED4	ED5	ExtLow	ExtHigh	ED2	ED3	ED4	ED5	ExtHigh
Chla	–	–	–	–	–	–	11	4	–	–	26
Corm	8	4	3	4	–	–	–	–	–	–	–
HabAv	na	na	na	na	na	na	20	13	16	13	–
HabPe	16	–	–	–	–	–	na	na	na	na	na
HabSt	2	7	2	–	–	–	–	–	–	4	41
Jet	–	–	–	–	–	–	14	13	5	6	–
Lat	11	2	–	–	–	–	8	5	4	1	–
MD	< 0.1	11	–	–	–	–	5	9	16	13	–
Ntot	3	–	–	–	–	–	12	–	7	–	29
Ptot	15	6	5	8	–	–	20	5	4	4	–
Sal	–	–	–	–	16	–	2	1	1	4	–
SDist	11	1	–	–	16	–	–	–	–	–	59
SWE	–	–	–	–	–	–	12	2	7	9	35
Temp	–	–	–	–	–	–	6	–	–	–	–
TempF	–	–	–	–	–	–	9	7	11	–	–
TrF	–	–	–	–	10	–	15	14	13	10	32
WExp	< 0.1	–	–	–	–	–	6	13	–	8	–
N feat.	9	6	3	2	3	0	13	11	10	10	6

3 ± 0.5 with the *Cyprinids indicator*, and that of summer temperature (*Temp*) between 16.5 and 17.0 °C.

In some cases the IEMD algorithm only separated extremely low or high values from the rest of the data points. For example, predation by cormorants (*Corm*) in the *Perch indicator* models and availability of stickleback habitats (*HabSt*) for both *Perch* and *Cyprinids indicators* were typically discretized into two bins, one including very low values and one with all the other data (Table A2; Table A3). In the ExtLow model of the *Perch indicator*, extremely high salinities (*Sal* > 7.2) and low water transparencies (*TrF* < 1.4 m) were found informative when predicting low values of the indicator.

3.3. Performance and functioning of the TAN models

For both indicators, the extreme TAN models (ExtLow for *Perch indicator* and ExtHigh for *Cyprinids indicator*) had the overall highest performance in terms of error rates (ER1 and ER2, Table 3). In general, the error rates were smaller with lower number of bins in the class variable (indicator). The outcomes are likely a result of the probability of correct prediction being higher when the number of bins is lower and their ranges, correspondingly, wider. The AUC metric indicated the best performance for models of the *Perch indicator* with fewer classes. The models of the *Cyprinids indicator* performed approximately equally according to the AUC metric (Table 3).

The ER1 error metric was equal to or smaller than the ER2 values in all cases, as ER1 predicts based on data that have already been used in the model parameterization phase, while ER2 predicts "new" data that the model has not yet observed. Also the difference between ER1 and ER2 decreased with the number of bins in the class variable (Table 3). For the ExtLow model of the *Perch indicator*, no difference was observed between ER1 and ER2. For the *Cyprinids indicator*, there was generally no differences between ER1 and ER2, except for a small difference in the ED5 model.

Table 3
Performance metrics of the alternative TAN-models for the Perch and Cyprinids indicators. For the explanation of the metrics used, see chapter 2.2.2. na = not applicable.

	Perch			Cyprinids		
	ER1	ER2	AUC	ER1	ER2	AUC
ED2	24.19	25.27	0.85	26.88	26.88	0.85
ED3	45.70	46.77	0.85	35.48	35.48	0.84
ED4	52.69	59.14	0.75	44.62	44.62	0.90
ED5	68.28	74.19	0.64	51.61	52.69	0.85
ExtLow	12.90	12.90	0.79	na	na	na
ExtHigh	na	na	na	5.38	5.38	0.87

Table 2 shows the entropy reduction of the class variables (indicators) per feature and model variant, averaged over all the alternative states of the feature, and when no other features are observed. Many of the features that were included in the IEMD phase, when modelled together, showed low informative significance for the class variable. Nevertheless, the availability of coastal fish habitats (*HabAv*) was among the most informative features in all the ED models of the *Cyprinids indicator*, the same applying to total phosphorus concentration (*Ptot*) in almost all of the *Perch indicator* models. In the extreme models of both indicators, all the included variables showed high informative significance, especially in the ExtHigh model of the *Cyprinids indicator*.

3.4. Example: Studying interactions among variables with a Bayesian network classifier

One way to use the Bayesian network classifiers resulting from the presented approach for diagnostic purposes is demonstrated here and in Fig. 2a–d. For this, we use the ExtHigh TAN model of the *Cyprinids indicator* to study dependencies between the variables. Fig. 2a shows the model in the state, where no observations on any variables are made. Fig. 2b represents the situation where the value of the class variable *Cyprinids indicator* (*Cypr*) is observed as "extremely high" (CPUE 30–50, see Fig. 1b) – a state typically associated with a high trophic (eutrophic) state. The probability of observing this state is 10.5%. The probability distributions of the features are updated accordingly. Distance to open sea (*SDist*) is expected to be in the higher class (51448–55399 m) with 69.9% probability, for example. Fig. 2c shows how the probabilities update when total nitrogen concentration (*Ntot*) is observed in its lower class (probability of occurrence being 82.7%). This observation decreases the probability of *Cypr* being in the "extremely high" class from 10.5% (Fig. 2a) to 3.3% (Fig. 2c). When *Ntot* is observed in its higher class (Fig. 2d), the uncertainty about the value of *Cypr* is much higher, as the probability mass is more evenly distributed among the possible outcomes.

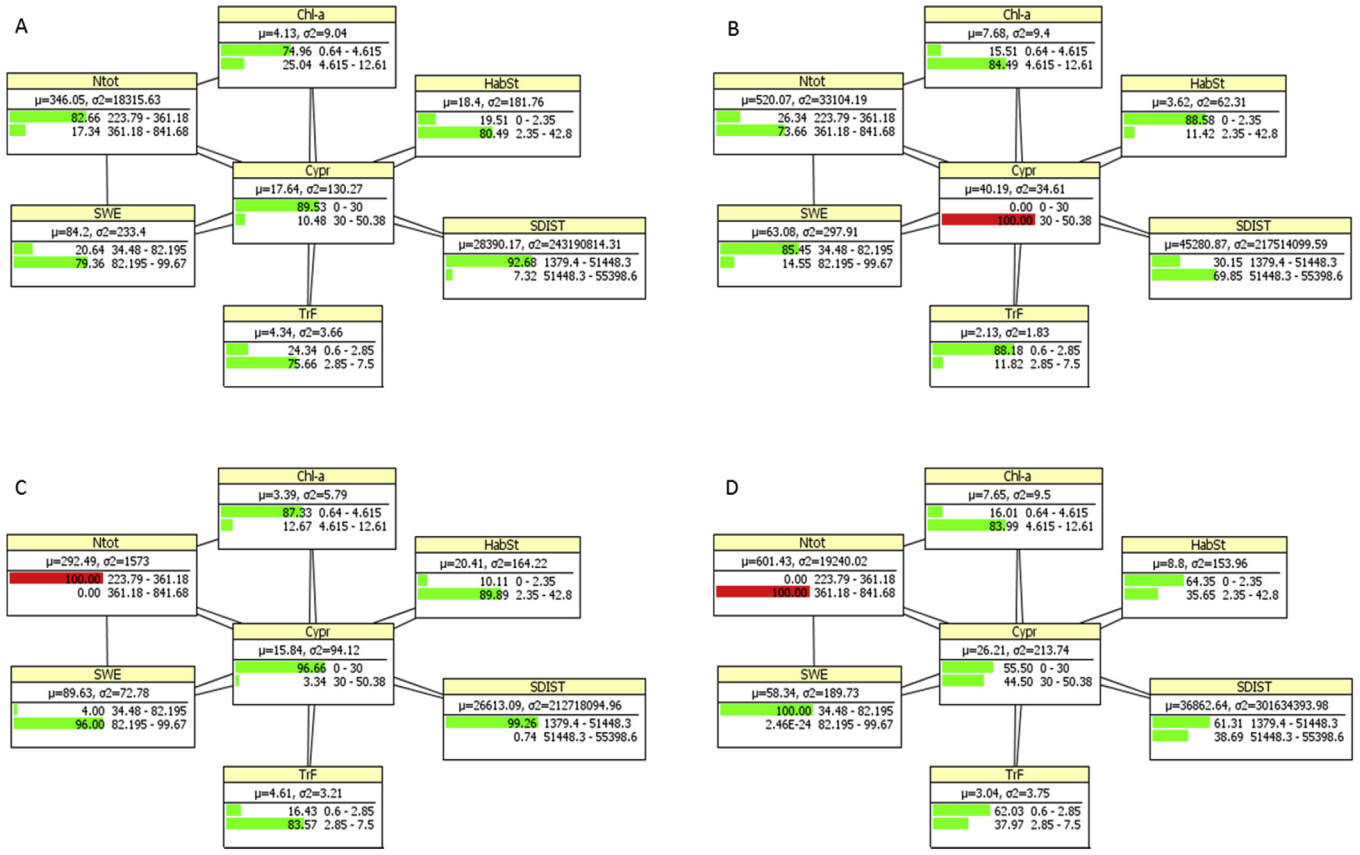

Fig. 2. Given updated information about the state of the Cyprinids indicator (B) or nitrogen concentration (C and D), the ExtHigh TAN model is used for reasoning about the likely state of the rest of the variables. The states with red bars are set to be known (P = 100%). Each observation in the model updates the probability distributions of the other variables according to the probabilistic dependencies learned from the data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

We tested whether machine learning based semi-supervised techniques could be useful for assessing the relationships between multiple environmental factors and ecological indicators. Such improved understanding could help societies in planning management strategies that most likely lead to reaching and maintaining good status of the ecosystem. The applied examples show that the method tested can be helpful in assessing the factors potentially influencing the status of the indicators, and what the relevant threshold values of these variables could be. This information could in turn be used in environmental management to demonstrate at which threshold value the state of an indicator is likely to respond to a pressure or a combination of pressures. In contrast to many multivariate statistical methodologies, the Bayesian classifiers can handle missing data and data of varying types, from fully quantitative to presence-absence, in the same analysis (e.g. Barber, 2012).

The example cases of this study highlight some interesting characteristics of the approach. The outcome of the analyses, starting from the model framing and structure, is strongly dependent on the number and location of the cut points chosen for the class variable. High dependency of the outcome on the discretization has been identified also in earlier studies, and even seen as a disadvantage proving the lack of robustness of the approach (Nojavan et al., 2017). The present analyses do, however, indicate that the level of this dependency is case specific (e.g. differences between the two example indicators), providing more information about the robustness of a particular case than the approach itself. Additionally, the fact that the outcome depends on the discretization can be extremely useful, as it can help in finding out the environmental factors (features) that are most relevant for the particular interesting cut points of the indicator (class variable). This is demonstrated in the current study through the analysis of the equally discretized versus the extremely high and low discretized indicator variables (Table 2). This analysis shows that some features are not significant for the equally discretized variable, but become relevant contributors for reaching the extremely high or extremely low values.

These findings, however, highlight how the discretization of the class variable should strictly reflect the formulation of the research question to which the model is intended to answer. In cases where clear pre-defined cut points of the class variable cannot be identified prior to the analysis, we recommend conducting several analyses using a relevant set of alternative versions of discretization. In all, the outcomes should be routinely interpreted with domain experts having deep understanding about the underlying environmental system (see also Fernandes et al., 2012). If the results are not logical in the sense that they do not follow a clear pattern and/or follow ecological theory, they must be treated with caution. Deviating patterns may nevertheless be used to improve current theories and create hypothesis for further testing.

Different steps of the presented machine learning approach provided different types of relevant information about the environmental factors potentially explaining the states of the fish community indicators. The search for informative discretization of the environmental factors over the alternatively discretized *Perch* and *Cyprinid indicators* helped us to identify the environmental factors having significant covariation with the indicator, given its cut points of interest. The number and location of the cut points identified for the environmental factors provide additional information about the resolution of the covariance and the change points in the data, respectively. When applied with varying discretization of an indicator and over different

indicators, the approach might also be valuable for finding potentially universal and robust ecological change points across ecosystem components and species in a system.

When a Bayesian network classifier is learned from data, the outcome represents the joint distribution over the whole set of variables (e.g. Korb and Nicholson, 2010). The resulting model represents the joint covariance of the variables, thus providing information about their likely co-occurrence, which can be studied for example as presented in Section 3.4. The entropy reduction analysis (as presented in Table 2) can be used to study the mutual informative significance of the features (here environmental factors), when their joint covariation with the class variable (indicator) and with each other are acknowledged. As demonstrated by the results, features might sometimes have significant covariation with the class variable alone, thus being picked and discretized by the IEMD, but might still not be very informative predictors after the model is learned. It is worth noting, however, that the entropy changes resulting from observing different states of the feature may be opposing; one state reducing and another increasing the entropy, and this may in some cases be the reason behind the low average entropy reduction. Thus we conclude that by studying the informative significance of the features over differently discretized models, it might be possible to identify features that are most robust predictors of the class variable. The average entropy reduction as the measure does not tell the whole story, however, and to reach full understanding on the informative value of a feature, the direction and strength of the entropy change produced by observing each state of the feature need to be tested separately.

The predictive performance of the models (Table 3) indicates how strongly the state of the fish stocks (as shown by the indicators) depends on the explanatory environmental factors included in the model. If we then know the values of the explanatory environmental factors that are of remarkable influence for the indicator value in the model, it might be possible to predict the indicator value in data-poor cases as shown in Fig. 2C and D. Such information may be very useful in supporting judgement of the status for areas that are poorly supported by direct measurements of the indicator. If the predictive performance of the model is shown to be high and the status of the indicator is predicted with high probability based on the known environmental conditions in the area, the sampling effort can be spared with only small risk of error. It is noteworthy, though, that when a purely data-based statistical model is used for evaluating a specific case (a sub-domain of the model; a certain combination of location, salinity and nutrient status, for example), the data pool underlying the obtained probability distributions decreases. This is something to be monitored during the analysis, as in an extreme case, the prediction may be based on only anecdotal evidence.

In the presented case example our major aim was to evaluate the interdependencies between the environmental variables and indicator variables based on a pre-defined data set, available for environmental managers in the area. If the intention is to extrapolate the approach in time or space, e.g. to predict the likely status of an indicator for new areas, the predictive capacity naturally depends on whether the observed dependencies hold in this new environment or not, i.e. whether the system is sufficiently similar to the one where the data originates from. If they do not seem to hold, the whole model teaching process, including the variable discretization, should be done based on data from a system with higher level of similarity.

In our test dataset, the robustness of the results across the different model variants varied markedly between the two indicators. In the *Perch indicator* models, there were substantial differences across the alternative models with differently discretized class variable with respect to which features were informative for predicting the indicator value (Table 2). The number of features picked by the discretization algorithm was also lower when the number of bins in the indicator was higher, reflecting that for this indicator most of the environmental variables were informative only on a rough scale. As a contrast, the *Cyprinids indicator* models were in general more similar across discretization alternatives, suggesting a higher robustness of this indicator in relation to the pressure variables.

The most important results regarding the datasets evaluated in this study can be summarized as follows: the *Cyprinids indicator* data can be more accurately and robustly predicted based on the environmental factors included in our study. This is shown by the weaker growth of the error rates along the increasing number of bins in the class variable, and by the higher AUC values compared to the *Perch indicator* (Table 3). Also the nearly non-existent differences between ER1 and ER2 indicate that the *Cyprinids indicator* models are likely to perform well also when predicting the state “out of the box”.

Based on this study, we believe that the presented approach can provide useful insights to the single and joint dependencies present in environmental datasets, acknowledging the related uncertainty, something that is often overlooked in ecological studies despite that they are crucial for implementing meaningful and cost-efficient management actions. It can potentially improve our understanding of cases, where several environmental factors are simultaneously acting on some target variable of interest. The general analysis protocol is provided in the support material (Table A6 and Fig. A2).

Declarations of interest

None.

Acknowledgements

This study was primarily part of the project “Statusklassning inom MSFD – kustfiskexemplet” (NV-08613-13), the funding being granted to Jens Olson by the Swedish EPA. The work was partly done under the projects WISE[1] (A.L., decision no. 312625), BONUS BLUEWEBS[2] (L.U.) and BONUS FUMARI[3] (L.U.), funded by the Academy of Finland[1,2,3] and EU BONUS (Art 185) programme[2,3]. We are grateful to Göran Sundblad for his help in producing the estimates of habitat availability as used in this study, and to all staff at the Department of Aquatic Resources of the Swedish University of Agricultural Sciences, who collected the data that served as the basis for this study.

Appendix A. Supplementary data

In the Appendix more information about the data used in the analyses is provided, including information about the fish monitoring areas and discretization (cut points) of the variables in different models. In addition the performance metrics comparison for different classifiers tested (in addition to TAN) and general analysis protocol are provided. Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2018.12.053.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. Proceedings of the 2nd international symposium on information, bn petrow, f. Czaki. Akademiai Kiado, Budapest.

Ashari, A., Paryudi, I., Tjoa, A.M., 2013. Performance comparison between naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. Int. J. Adv. Comput. Sci. Appl. 4 (11), 33–39.

Barber, D., 2012. Bayesian Reasoning and Machine Learning. Cambridge University Press.

Bergström, L., Bergström, U., Olsson, J., Carstensen, J., 2016. Coastal fish indicators response to natural and anthropogenic drivers–variability at temporal and different spatial scales. Estuar. Coast. Shelf Sci. 183, 62–72.

Bergström, U., Olsson, J., Casini, M., Eriksson, B.K., Fredriksson, R., Wennhage, H., Appelberg, M., 2015. Stickleback increase in the Baltic Sea – a thorny issue for coastal predatory fish. Estuar. Coastal Shelf Sci. 163, 134–142.

Birk, S., Bonne, W., Borja, A., Brucet, S., Courrat, A., Poikane, S., Solimini, A., van de Bund, W., Zampoukas, N., Hering, D., 2012. Three hundred ways to assess Europe’s surface waters: an almost complete overview of biological methods to implement the water framework directive. Ecol. Ind. 18, 31–41.

Chow, C.K., Liu, C.N., 1968. Approximating discrete probability distributions with dependence trees. IEEE Trans. Inf. Theory 14 (3), 462–467.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Stat. Soc.: Ser. B (Methodol.) 39 (1), 1–38.

EC, 2008. Directive 2008/56/EC of the European Parliament and of the Council of 17 June 2008 establishing a framework for community action in the field of marine environmental policy. Off. J. Eur. Commun. L164, 19–40.
Farnaz, N.A., Song, S.Q., Stow, C.A., 2017. Comparative analysis of discretization methods in Bayesian networks. Environ. Modell. Software 87, 64–71.
Fayyad, Usama M., Irani, Keki B., 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the International Joint Conference on Uncertainty in AI (Q334 .I571 1993), pp. 1022–1027. http://yaroslavvb.com/papers/fayyad-discretization.pdf.
Fernandes, J.A., Kauppila, P., Uusitalo, L., Fleming-Lehtinen, V., Kuikka, S., Pitkänen, H., 2012. Evaluation of reaching the targets of the water framework directive in the Gulf of Finland. Environ. Sci. Technol. 46 (15), 8220–8228.
Flach, P., 2012. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press.
Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifiers. Machine Learn. 29, 131–163.
Hattam, C., Atkins, J.P., Beaumont, N., Börger, T., Böhnke-Henrichs, A., Burdon, D., de Groot, R., Hoefnagel, E., Nunes, P.A.L.D., Piwowarczyk, J., Sastre, S., Austen, M.C., 2015. Marine ecosystem services: linking indicators to their classification. Ecol. Ind. 49, 61–75.
Hand, D.J., Yu, K., 2001. Idiot's Bayes – not so stupid after all? Int. Statist. Rev. 69 (3), 385–398.
HELCOM, 2007. HELCOM Baltic Sea Action Plan (BSAP). HELCOM Ministerial Meeting. Adopted in Krakow, Poland, 15 November 2007. Online. [2018-11-23], [http://www.helcom.fi/Documents/Baltic%20sea%20action%20plan/BSAP_Final.pdf].
HELCOM, 2018a. Abundance of key coastal fish species HELCOM core indicator 2018. HELCOM core indicator report. Available at: http://www.helcom.fi/baltic-sea-trends/holistic-assessments/state-of-the-baltic-sea-2018/reports-and-materials (accessed 2018-11-23).
HELCOM, 2018b. Abundance of coastal fish key functional groups HELCOM core indicator 2018. HELCOM core indicator report. Available at: http://www.helcom.fi/baltic-sea-trends/holistic-assessments/state-of-the-baltic-sea-2018/reports-and-materials (accessed 2018-11-23).
HELCOM, 2018c. State of the Baltic Sea – second HELCOM holistic assessment 2011–2016. Baltic Sea Environ. Proc. 155 ISSN: 0357-2994.
HELCOM, 2018d. Status of coastal fish communities in the Baltic Sea during 2011–2016 – the third thematic assessment. Baltic Sea Environ. Proc. 161 ISSN: 0357-2994.
Isæus, M., 2004. Factors Structuring *Fucus* Communities at Open and Complex Coastlines in the Baltic Sea. Doctoral thesis. Stockholm University, Stockholm.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer, New York (Vol. 112).
Korb, K., Nicholson, A., 2010. Bayesian Artificial Intelligence. Chapman & Hall/CRC Press ISBN: 9781439815915, second ed.
Kuncheva, L.I., 2006. On the optimality of Naive Bayes with dependent binary features. Pattern Recogn. Lett. 27 (7), 830–837.
Kupschus, S., Schratzberger, M., Righton, D., 2016. Practical implementation of ecosystem monitoring for the ecosystem approach to management. J. Appl. Ecol. 53, 1236–1247.
Landgrebe, T., Duin, R., 2006, November. A simplified extension of the area under the ROC to the multiclass domain. In: Seventeenth annual symposium of the pattern recognition association of South Africa, pp. 241–245.
Large, S.I., Fay, G., Friedland, K.D., Link, J.S., 2013. Defining trends and thresholds in responses of ecological indicators to fishing and environmental pressures. ICES J. Mar. Sci. 70, 755–767.
Lauritzen, S.L., 1995. The EM algorithm for graphical association models with missing data. Comput. Stat. Data Anal. 19 (2), 191–201.
Madsen, A.L., Jensen, F., Kjærulff, U.B., Lang, M., 2005. The Hugin tool for probabilistic graphical models. Int. J. Artif. Intell. Tools 14 (3), 507–543.
Murphy, K., 2012. Machine Learning. A Probabilistic Perspective. Massachussetts Institute of Technology.
Nojavan, F.A., Qian, S.S., Stow, C.A., 2017. Comparative analysis of discretization methods in Bayesian networks. Environ. Modell. Software 87, 64–71.
Östman, Ö., Eklöf, J., Eriksson, B.K., Olsson, J., Moksnes, P.O., Bergström, U., 2016. Top-down control as important as nutrient enrichment for eutrophication effects in North Atlantic coastal ecosystems. J. Appl. Ecol. 53, 1138–1147.
Östman, Ö., Lingman, A., Bergström, L., Olsson, J., 2017. Temporal development and spatial scale of coastal fish indicators in reference sites in coastal ecosystems: hydroclimate and anthropogenic drivers. J. Appl. Ecol. 54, 557–566.
Rice, J.C., Rochet, M.-J., 2005. A framework for selecting a suite of indicators for fisheries management. ICES J. Mar. Sci. 62 (3), 516–527.
Schwarz, G., 1978. Estimating the dimension of a model. Annals Stat. 6 (2), 461–464.
SMHI, 2016. Vattenweb-database of Swedish Meteorological and Hydrological Institute. Online. [http://vattenweb.smhi.se/].
Spirtes, P., Glymour, C., Scheines, R., 2000. Causation, Prediction, and Search. Adaptive Computation and Machine Learning, second ed. MIT Press.
Sundblad, G., Bergström, U., Sandström, A., Eklöv, P., 2013. Nursery habitat availability limits adult stock sizes of predatory coastal fish. ICES J. Mar. Sci. 71 (3), 672–680.
Sundblad, G., Bekkby, T., Isæus, M., Nikolopoulos, A., Norderhaug, K.M., Rinde, E., 2014. Comparing the ecological relevance of four wave exposure models. Estuar. Coast. Shelf Sci. 140, 7–13.
Sundblad, G., Bergström, U., 2014. Shoreline development and degradation of coastal fish reproduction habitats. Ambio 43, 1020–1028.
Tam, J.C., Link, J.S., Rossberg, A.G., Rogers, S.I., Levin, P.S., et al., 2017. Towards ecosystem-based management: identifying operational food-web indicators for marine ecosystems. ICES J. Mar. Sci. 74, 2040–2052.
Teixeira, H., Berg, T., Uusitalo, L., Fürhaupter, K., Heiskanen, A.-S., Mazik, K., Lynam, C.P., Neville, S., Rodriguez, J.G., Papadopoulou, N., Moncheva, S., Churilova, T., Kryvenko, O., Krause-Jensen, D., Zaiko, A., Veríssimo, H., Pantazi, M., Carvalho, S., Patrício, J., Uyarra, M.C., Borja, À., 2016. A catalogue of marine biodiversity indicators. Front. Marine Sci. 3, 207 p.
Törnqvist, O., Engdahl, A., 2010. Kartering och analys av fysiska påverkansfaktorer i marin miljö. Swedish Environmental Protection Agency, Report 6376, Stockholm, Sweden, 79 pp (in Swedish, English summary).
Uusitalo, L., Korpinen, S., Andersen, J.H., Niiranen, S., Valanko, S., Heiskanen, A.S., Dickey-Collas, M., 2016. Exploring methods for predicting multiple pressures on ecosystem recovery: a case study on marine eutrophication and fisheries. Cont. Shelf Res. 121, 48–60.
Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental modelling. Ecol. Model. 203 (3–4), 312–318.
Zhang, H., 2004. The optimality of naive Bayes. Proceedings of the 17th International FLAIRS Conference. AAAI Press.
Zheng, F., Webb, G.I., 2010. Tree augmented naive bayes. In: Sammut, C., Webb, G.I. (Eds.), Encyclopedia of Machine Learning. Springer, US, Boston, MA.