Norwegian Financial Mechanism and Republic of Lithuania

Coastal Research and Planning Institute, Klaipėda University

Project: A system for the sustainable management of Lithuanian marine resources using novel surveillance, modeling tools and ecosystem approach

Technical Report No. 6

SPATIAL DISTRIBUTION OF FISH FEEDING GROUNDS IN THE LITHUANIAN EXCLUSIVE ECONOMIC ZONE

Project indicator:
1. Documented assessment of fish feeding ground quality

Prepared by: A. Šiaulys
Contributors: M. Bučas,
D. Daunys

Coastal Research and Planning Institute, Klaipeda University

March 2011, Klaipėda

TABLE OF CONTENTS

INTRODUCTION

One of the most important ecosystem service provided by the seafloor is the feeding grounds for many fish species. Benthic macrofauna is widely known as being an important food sources for higher trophic levels in marine ecosystems. Large Baltic fish species such as cod or flounder apart from small fish can intensively feed on wide spectra of benthic invertebrates such as isopods *Saduria entomon*, bivalves *Macoma balthica, Mytilus edulis, Mya arenaria* and even relatively small polychaete worms and gammarids.

In order to obtain knowledge on the spatial distribution of benthophagous fish feeding grounds information on spatial distribution of prey items is needed, however the data is usually point-based. To convert point-based data of biomass of fish prey items to continuous layer of biomass distribution modelling can be performed. Overlaying spatial distribution layers of fish prey items allows developing distribution maps of food resources for a particular fish species. These maps can be a very useful tool for spatial planning and decision making.

1. MATERIALS AND METHODS

1.1 Identification of fish diet composition

Feeding grounds were analysed for three common fish species in the exclusive Lithuanian economic zone (LEZ): Baltic cod (*Gadus morhua callarias L.*), flounder (*Platichthys flessus L.*) and eelpout (*Zoarces viviparus L.*). Baltic cod was selected because of its importance for both commercial and recreational fisheries. Flounder is very common in the LEZ and is found in high abundances in a wide depth range during all seasons. Eelpout is found only in coastal and transitional waters and is recommended for protection in Lithuania (Repečka, 2003).

To assess diets of these fish species 1425 digestive tracts (empty tracts are not included) were analysed (Table 1).

Table 1. The depth range of sampled fish and the number of digestive tracts analysed

	Baltic cod	Flounder	Eelpout
Depth range of fish sampling, m	12-79	6-79	6-30
No. of digestive tracts analysed	300	1000	125

Table 2 shows the occurrence and importance of prey items. Occurrence describes how often a particular prey item is found in the digestive tract and corresponds how frequently fish feed on it. "*High*" occurrence means that particular benthic animal is found almost in every tract and "*low*" means it is found rarely, but not accidentally. Importance describes how much a particular prey item can contribute to the total content in the digestive tract. "*High*" importance means that almost whole digestive tract can be filled with a particular prey species; "*low*" means that a particular item is only a small addition to the whole tract content.

Table 2. Occurrence in digestive tracts (upper entry) and importance (lower entry) of prey items for cod, flounder and eelpout. Empty cells indicate that fish do not prey on the particular item.

Prey items	Fish species: Baltic cod	Flounder	Eelpout
Gammarus spp.	High High	High High	High Moderate
Halicryptus spinulosus		Moderate Moderate	
Hediste diversicolor	Moderate Low	Moderate Low	Moderate Low
Macoma balthica		High High	Moderate Moderate
Marenzelleria neglecta	Low Low	Low Low	Low Low
Mya arenaria		Low Low	
Mytilus edulis		Moderate Moderate	Low Low
Saduria entomon	High High	Moderate Low	Moderate Low

1.2 Modelling of fish prey items

1.2.1 Modelling technique

For the prediction of the biomass distribution of zoobenthos species Random forests (Breiman, 2001) statistical model implemented in the "randomForest 4.6-2" package (Liaw, Wiener, 2002) within the R environment was chosen. Intermediate results from the project PREHAB (Spatial PREdiction of benthic HABitats in the Baltic Sea: incorporating anthropogenic pressures and economic evaluation) showed that RF performed very well in comparison to other techniques: GAM (Generalized Additive Models), MARS (Multivariate Adaptive Regression Splines), KED (Kriging with External Drift) and MaxEnt (Maximum-Entropy modelling). Also several studies have shown that RF models often reach top predictive performances compared to other methodologies (Kuhn et al., 2008; Vincenzi et al., 2011).

As described by Prasad et al. (2006) Random Forests (RF) is a new entry to the field of datamining and is designed to produce accurate predictions that do not overfit the data. RF is similar to Bagging Trees (BT) in that bootstrap samples are drawn to construct multiple trees; the difference is that the each tree is grown with a randomized subset of predictors, hence the name ''random'' forests. A large number of trees (500 to 2000) are grown, hence a ''forest'' of trees. The number of predictors used to find the best split at each node is a randomly chosen subset of the total number of predictors. As with BT, the trees are grown to maximum size without pruning, and aggregation is by averaging the trees. Out-of-bag samples can be used to calculate an unbiased error rate and variable importance, eliminating the need for a test set or cross-validation. Because a large number of trees are grown, there is limited generalization error (that is, the true error of the population as opposed to the training error only), which means that no overfitting is possible, a very useful feature for prediction.

1.2.2 RF model procedure

- Creating a correlation matrix for all predictors. If a correlation coefficient is higher than 0.7 or VIF (variance inflation factors) value is higher than 3, those predictors are not used for model build-up.
- Splitting biomass data into two datasets: train data (70% of all data) for building a model and test data (rest 30%) for validating the model. In order to avoid uneven distribution of zero values the split is made semi-randomly. That means that all sites are chosen randomly, but ensuring that sites with zero values would distribute with rate 70/30 % in train/test datasets.
- Selecting of parameters for RF. Three main parameters should be defined for RF models (Vincenzi et al., 2011): number of trees to grow (*ntree* or *jbt*), number of variables randomly selected at each node (*mtry*) and minimum node size (*nodsize* or *ndsize*). *ntree* was set to 1000, *mtry* and *ndsize* were set to default values 2.[3] and 5 respectively.
- Running the model.
- Getting the importance of variables. Mean Decrease Accuracy (%IncMSE) and Mean Decrease Gini (IncNodePurity) are calculated to assess the importance of every environmental factor for the response variable. Partial plots are drown reflecting dependence of a particular factor and the response variable.
- Internal and external model validation. In model validation phase observed values (from sample data) are compared against predicted values produced by the model. During internal validation predicted values are compared against internal data (train dataset), which was used for the model build-up thus meaning little about model performance. During external validation predicted values are compared against external data (test dataset) thus revealing true model performance. Three estimates are calculated: RMSE (root mean square error), MAD (mean absolute deviation) and r_s (correlation between observed and predicted values).
- Exporting to GIS. Finally, predictions are made for the whole research area in 100x100 meter grid and together with coordinates are transcribed to DBF file which can be easily used with most GIS programs.

1.2.3 Environmental predictors

Seven environmental factors (predictors) were used for the prediction of biomass of zoobenthos species. Numerical predictors were salinity, near-bottom oxygen, near-bottom current velocity, orbital velocity, categorical – sediment types, areas of presence or absence of thermocline and areas above and below halocline (Figure 1). Some parameters like depth, slope and Secchi depth were intentionally removed from models because of their high correlations with other predictors thus violating modelling conditions.

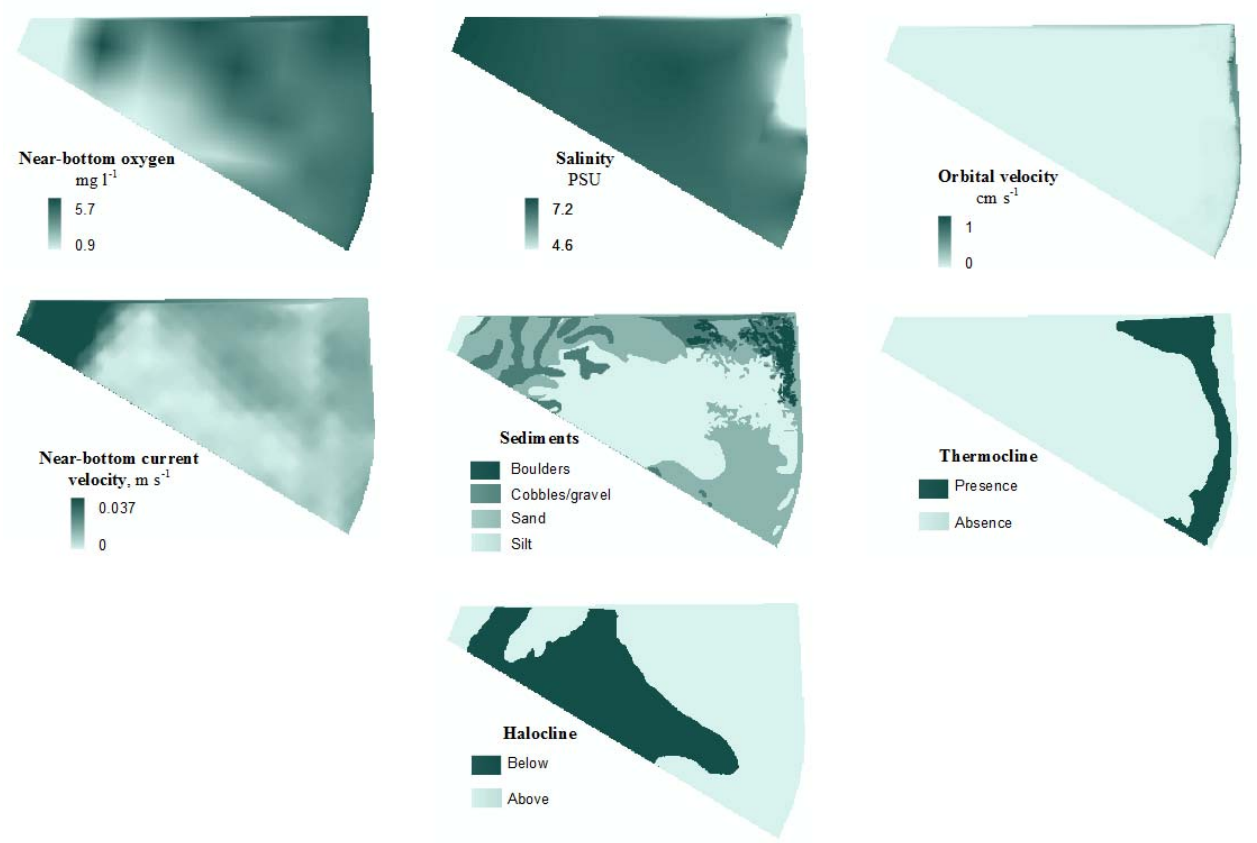

Figure 1. Environmental parameters used for modelling and prediction of biomass of macrozoobenthos

The strength of predictors to a response variable are expressed as percents of Mean Decrease Accuracy (%IncMSE). As described by Kuhn et al. (2008) %IncMSE is constructed by permuting the values of each variable of the test set, recording the prediction and comparing it with the unpermuted test set prediction of the variable (normalised by the standard error). For classification (presence-absence data), it is the increase in the percentage of times a test set tuple is misclassified when the variable is permuted. For regression (biomass data), it is the average increase in squared residuals of the test set when the variable is permuted. A higher %IncMSE value represents a higher variable importance.

Partial plot function (Friedman, 2001) was used to plot the dependence between a particular predictor and the response variable.

1.2.4 Field data on prey items

Overall 224 sampling sites and 640 benthic samples taken from period of 1998-2010 y. were used for modelling. 165 samples were taken with a Van-veen grab, 59 – by SCUBA divers. Samples were taken and treated following standard guidelines for bottom macrofauna sampling (HELCOM, 1988).

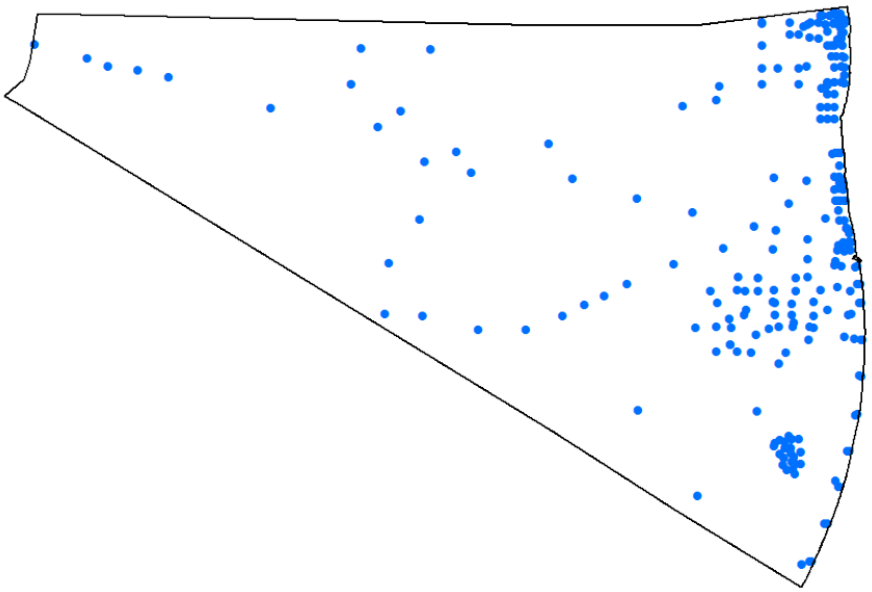

Figure 2. The distribution of sampling sites in the research area

1.2.5 Validation of models

Several estimates are calculated for model validation: (1) RMSE – root mean squared error, (2) MAD – mean absolute deviation, CV(RMSE) – coefficient of variation of the RMSE, CV(MAD) – coefficient of variation of the MAD, r_s – correlation between observed (y_t) and predicted ($\hat{y}_t$) values.

$$RMSE = \sqrt{n^{-1}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \quad (1)$$

$$MAD = n^{-1}\sum_{t=1}^{n}|y_t - \hat{y}_t| \quad (2)$$

$$CV(MAD) = \frac{MAD_{ext.}}{\bar{y}} \times 100 \quad (3)$$

Both RMSE and MAD shows an average error made by the model, however MAD is slightly easier to interpret, thus it was selected for further estimation of model accuracy. If MAD is divided by $\bar{y}$ (an average of y_t) and expressed as percentages, we can estimate a relative difference between the average error and the average of observations. It is called coefficient of variation of MAD and abbreviated as (3) CV(MAD). The smaller CV values are, the higher accuracy of the model is. CV values are dimensionless and since there is a division from the average it is possible to compare these values among all models. Accuracy of model was calculated simply by dividing CV(MAD) from 100. Accuracy of 100% means that predictions are without errors (impossible to achieve), 0% means that prediction error is equal to the sample average.

For estimation of the relationship between observations and predictions Spearman's rank correlation was selected. Due to a relatively high number of zero values data cannot be described by Gaussian distribution.

1.3 Development of fish feeding ground maps

1.3.1 Creation of the fish feeding grounds maps

The output file of a model consists of predicted biomass values in 100x100 meter grid and is imported in ArcGIS 9.3.1 software. Using "Natural Neighbor" interpolation in "3D Analyst Tools" ArcToolbox the raster files of biomass distribution are produced (see chapter 4.2).
Using "Raster Calculator" tool in "Spatial Analyst Tools" ArcToolbox rasters of those prey items that a particular fish species feeds on are added up with different weights (Table 3). Weights are given according to *occurrence* and *importance* shown in table 2. Initial biomass values are multiplied by the weight value, that way feeding items which are more important are reflected better in the map of feeding grounds.

Table 3. Raster weights depending on occurrence and importance

Frequency/Part of diet	**Weight**
High/High	1
High/Moderate *or* Moderate/High	0.67
Moderate/Moderate	0.5
Moderate/Low *or* Low/Moderate	0.33
Low/Low	0.25

As different multipliers were used, biomass units (g m^{-2}) are no longer suitable. So weighted biomass was categorised into five types of importance: *very high*, *high*, *moderate*, *low* and *very low*, where *very high* importance indicates the highest weighted biomass and *very low* – lowest biomass.

1.3.2 Accuracy of maps

Three ranks of accuracy (high, moderate and low) are given for maps of seabed importance for fish feeding. Basically, accuracy reflects how well different intervals of predictors are justified by field data. The idea is that if the interval/category of a particular predictor is justified by relatively high number of samples, the accuracy in the area of this interval/category must be high. If there are only few samples in the interval/category – the accuracy must be low.

First of all the accuracy of separate prey item biomass maps are estimated. The first step is to calculate the number of samples per particular interval/category of predictors (Table 4). Since 171 samples were used for the model build up, 171 is the total point pool that is split between all intervals/categories of a single predictor. Then "Reclassify" tool in "3D Analyst Tools" ArcToolbox is used to reclassify predictor layer assigning these points for all intervals/categories. Using "Raster Calculator" tool in "Spatial Analyst Tools" ArcToolbox reclassified rasters of those prey items that a particular fish species feeds on are added up with different weights. Weights are assigned equal to %IncMSE values (Table 5). That way the accuracy of the most important predictor receives highest weight and minor predictors has only an insignificant impact on overall accuracy. Finally, 50% and 80% thresholds were selected for assigning *low* (<50%), *moderate* (50-80%) and *high* (>80%) accuracy.

Table 4.Number of field samples per category or interval of environmental predictor

	Category/interval	Halicryptus	Gammarus	Marenzelleria	Hediste	Mytilus	Mya	Macoma	Saduria
SEDIMENTS	boulders	21	24	24	22	23	21	22	23
	cobbles/gravel	16	12	16	14	14	18	15	16
	sand	96	95	88	99	83	88	96	96
	silt	38	40	43	36	45	44	38	36
THERMOCLINE	absence	140	142	137	145	131	131	136	135
	presence	31	29	34	26	34	40	35	36
HALOCLINE	absence	164	159	159	159	153	158	160	47
	presence	7	12	12	12	12	13	11	124
NEAR-BOTTOM CURRENT VELOCITY	0	26	33	33	33	33	38	36	35
	0-0.01	141	135	135	134	129	130	131	131
	0.01-0.02	1	0	1	1	1	1	1	1
	0.02-0.03	3	3	2	3	2	2	3	4
	>0.03	0	0	0	0	0	0	0	0
ORBITAL VELOCITY	0	14	13	17	15	32	16	11	11
	0-0.2	121	124	125	124	106	131	127	128
	0.2-0.4	21	19	15	18	14	15	20	20
	0.4-0.6	15	15	14	14	13	9	13	12
	>0.6	0	0	0	0	0	0	0	0
SALINITY	<4.8	0	0	0	0	0	0	0	0
	4.8-5.4	12	14	7	10	6	12	11	11
	5.4-6.0	19	20	15	18	15	17	19	16
	6.0-6.6	39	36	36	39	42	38	37	38
	6.6-7.2	101	101	113	104	102	104	104	106
	>7.2	0	0	0	0	0	0	0	0
NEAR-BOTTOM OXYGEN	<0.9	0	0	0	0	0	0	0	0
	0.9-2.06	3	2	2	2	1	2	2	3
	2.06-3.22	3	4	4	5	6	6	4	4
	3.22-4.38	19	27	24	26	22	24	26	22
	4.38-5.53	146	138	141	138	136	139	139	142
	>5.53	0	0	0	0	0	0	0	0

2. RESULTS AND DISCUSION

2.1 Performance of models

2.1.1 Predictors

Seven environmental predictors (Figure 1) were used in modelling of biomass distribution of zoobenthos (short titles used in models and some graphs are given in brackets): salinity (SALINITY), near-bottom oxygen (BOXYGEN), near-bottom current velocity (BCURRENT), orbital velocity (ORBITALBV), sediment types (SEDIMENTS), areas of presence or absence of thermocline (THERMOCLINE) and areas above and below halocline (HALOCLINE). Figure 3 schematically shows the importance of environmental variables for predictions of all zoobenthos species. The most important predictor were near-bottom oxygen level. Orbital velocity, salinity and sediments were also important, near-bottom current velocity were less important, halocline and thermocline had only a minor importance or no importance at all in some cases.

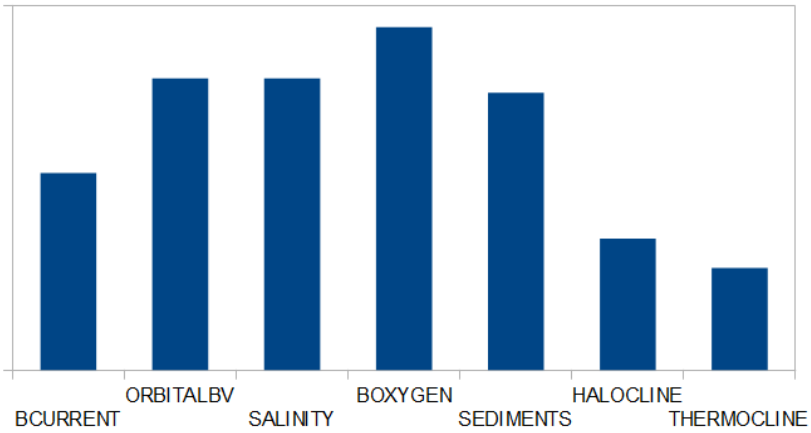

Figure 3. Importance of predictors for all response variables

Mean decrease accuracy was calculated for each predictor in order to evaluate their importance to the response variable (Table 5).

Table 5. Mean decrease accuracy (%IncMSE) of environmental predictors. Higher value indicates higher importance.

	Species							
Predictors	*Gammarus spp.*	*H.diversicolor*	*H.spinulosus*	*M.arenaria*	*M.balthica*	*M.edulis*	*M.neglecta*	*S.entomon*
BCURRENT	3.8	6.4	3.5	7.6	22.4	3.9	0.5	7.2
ORBITALBV	2.4	12.0	12.7	6.9	18.0	9.6	7.9	18.9
SALINITY	6.7	16.3	3.8	0.2	25.1	17.0	7.4	15.0
BOXYGEN	7.1	10.7	12.1	9.2	28.7	16.1	3.9	24.6
SEDIMENTS	9.3	3.8	7.7	0.7	22.2	34.8	4.7	10.1
HALOCLINE	0.4	3.2	5.5	4.6	-1.4	1.4	1.3	6.3
THERMOCLINE	1.8	2.7	0.4	-4.2	14.4	10.4	0.7	5.8

Near-bottom oxygen level was more important for deep living species like *Macoma balthica, Saduria entomon, Halicryptus spinulosus* (28.7, 12.1 and 24.6 %IncMSE respectively). For these species low oxygen conditions usually set the boundary how deep these animals can live.

Distribution of biomass of *Gammarus* genus and *Mytilus edulis* was mostly dependent on sediments (9.3 and 34.8 %IncMSE respectively). That was expected because these benthic animals are mostly associated to single type of substrate – hard bottom. On the contrary, some species like *S.entomon* and *Hediste diversicolor* are found in almost on every substrate, so naturally sediment type as a predictor were of minor importance. At the first glance it seems to be strange that a typical soft-bottom species *Mya arenaria* is not dependent of sediments at all in model. It's probably because juveniles of this mollusc can be found attached on red algae *Furcellaria lubricalis* or even directly on the hard substrate.

There is no strong salinity gradient in the modelled area except the northern part of coastal zone. This gradient is formed due to the fresh water outflow from the Curonian lagoon. The low salinity area seems to be attractive for deposit feeding polychaete worms *H.diversicolor* and *M.neglecta* (notice first peaks of partial plots in Figure 4).

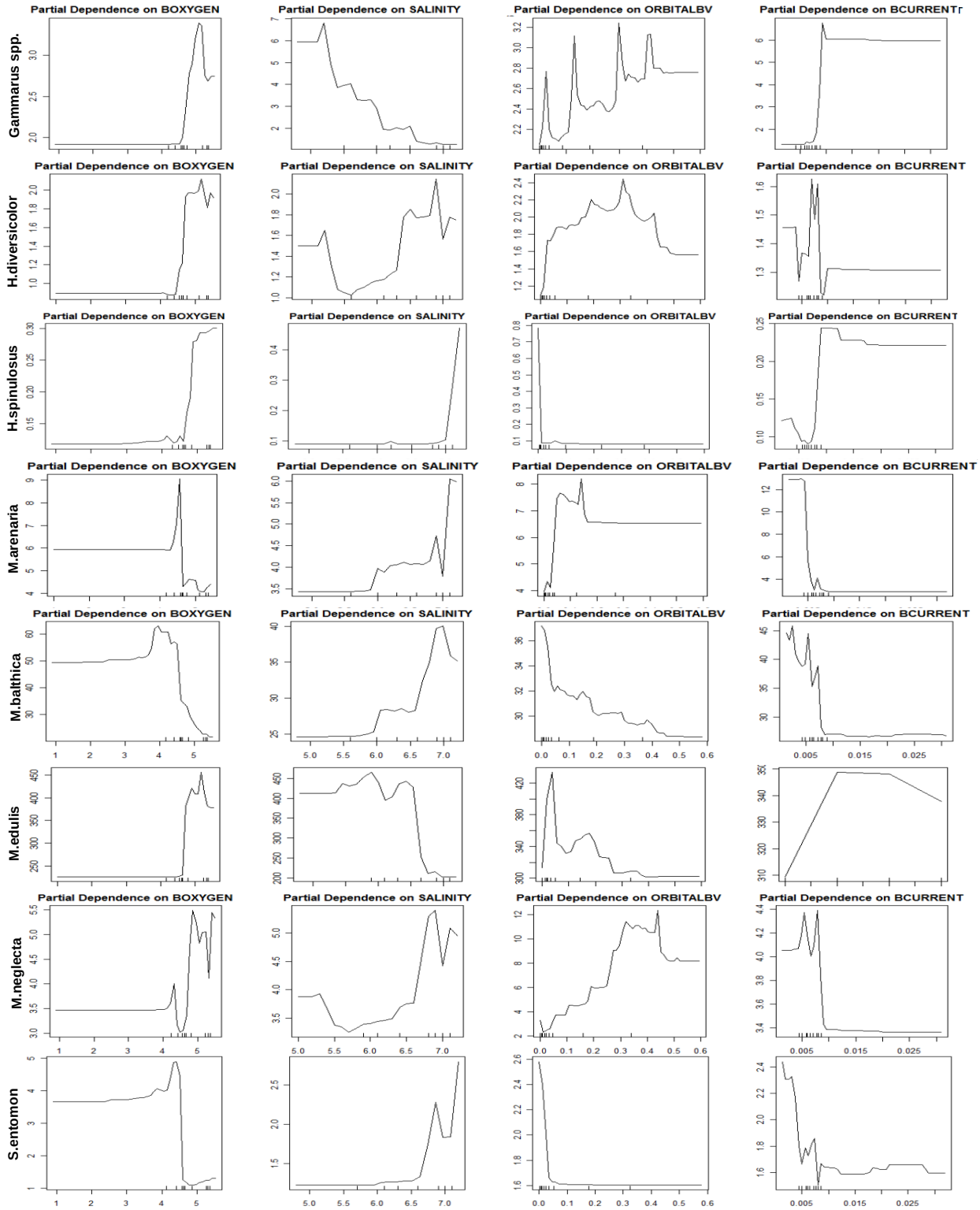

Figure 4.Partial plots between numerical environmental predictors and modelled species. X-axis: predictor values; Y-axis: biomass, g m^{-2}

Orbital velocity basically reflects wave exposure. Relatively high %IncMSE values for *H.spinulosus* and *S.entomon* (12.7 and 18.9 respectively) indicates high importance of this predictor for these species. Partial plots (Figure 4) explains it clearly, that these species are not to be found in exposed bottom (biomass curves dramatically decreases with even a slight increase of

orbital velocity). Partial plots of bivalves *M.balthica* and *M.edulis* reveal similar but not so dramatic tendencies, meaning that highest biomasses occur in energetically inactive areas, low biomass occurrences in moderately active areas and no presence in highly active areas. Some species like polychaetes and juveniles of *M.arenaria* can tolerate even high exposure. *Gammarus spp.* find shelter from waves in habitat forming species (red and green macroalgae, aggregations of blue mussels).

Thermocline and halocline were worst predictors almost for all models. This is probably because these environmental predictors were categorized into too large groups. As biomasses can variate even in small-scale, the large-scale predictors have minor importance for these models.

2.1.2 Validation of models

Model validation was separated into internal and external validation. During internal validation only train dataset is used both for the model build-up and calculation of prediction errors. Naturally, predictions errors in this case would be irrelevant. In external validation predictions based on the train dataset is compared with test dataset which wasn't used for model build-up. In this case validation results reveal true model prediction errors.

Output values of internal and external validation of all models are given in table 6. Highest prediction errors ($MAD_{ext.}$) were estimated for bivalves *M.edulis* and *M.balthica* (223.5 and 26.4 respectively, lowest for priapulid and *polychaete* worms *H.spinulosus* and *H.diversicolor* (0.1 and 1.4 respectively). This is expected because higher average or range values generate bigger $MAD_{ext.}$, so MAD says little about model performance in the context of other models if we do not take the average into account. For this reason CV(MAD) are calculated.

Table 6. Validation results of zoobenthos biomass models. In columns from left to right: species; average sample biomass and standard deviation; internal root mean squared error (RMSE); internal mean absolute deviation (MAD); correlation of observations and internal predictions; external RMSE; external MAD; correlation of observations and external predictions; coefficient of variation of MAD (CV(MAD))

	Mean biomass	**$RMSE_{int.}$**	**$MAD_{int.}$**	**$r_{s\ int.}$**	**$RMSE_{ext.}$**	**$MAD_{ext.}$**	**$r_{s\ ext.}$**	**CV(MAD)**
Gammarus spp.	7.8±16.3	4.4	1.6	0.51	10.6	2.6	0.48	33.2
H.diversicolor	2.0±3.2	1.5	0.8	0.74	2.6	1.4	0.57	71.2
H.spinulosus	0.3±1.1	0.4	0.1	0.57	0.4	0.1	0.46	38.1
M.arenaria	6.5±17.4	8.7	2.7	0.61	5.1	3.0	0.41	46.3
M.balthica	43.4±53.8	19.3	12.3	0.88	42.4	26.4	0.77	60.9
M.edulis	1385.4±1398.9	480.7	207.4	0.66	486.6	223.5	0.62	16.1
M.neglecta	3.8±9.4	6.1	2.6	0.74	4.0	2.7	0.34	70.1
S.entomon	5.6±5.8	1.8	0.8	0.76	3.7	1.8	0.76	32.6

If we put CV(MAD) values of all models together (Figure 5) it is possible to compare their errors/accuracies. The most accurate model was of *M.edulis.* This is probably because this mollusc strictly occurs only on hard substrate and usually in high biomass thus making an easier task for the predictive model. Models of *S.entomon*, *Gammarus spp*., *H.spinulosus* and *M.arenaria* were also quite accurate (accuracy >50%). The model of *M.balthica* was less accurate (<40%), the lowest accuracy was estimated for polychaete models (<30%) probably due to their patchy distribution.

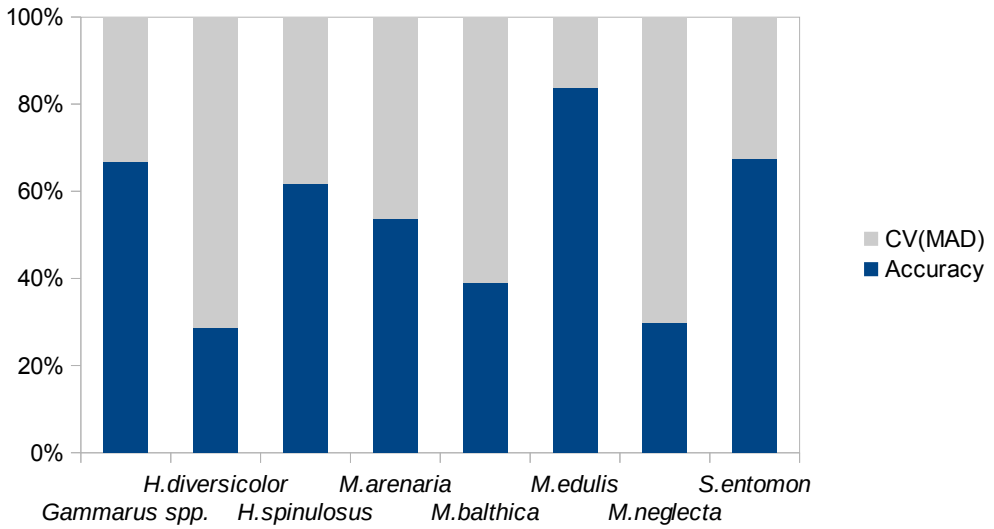

Figure 5. Accuracy and coefficients of variation of mean absolute deviation(CV(MAD)) for all models. Accuracy of 100% means that predictions are without errors (impossible to achieve), 0% means that prediction error is equal to the sample average.

Looking from a broader view overall accuracy of models were adequate, especially taking into account that biomass is hard to model (in comparison to presence-absence or abundance). None of models had errors higher than average of all samples so there should be enough confidence in maps of modelled biomasses.

2.2 Predicted biomass maps of fish prey items

2.2.1 Gammarus spp.

The prediction map of gammarids biomass (Figure 6) clearly indicates that this genus is mostly associated to hard bottoms.

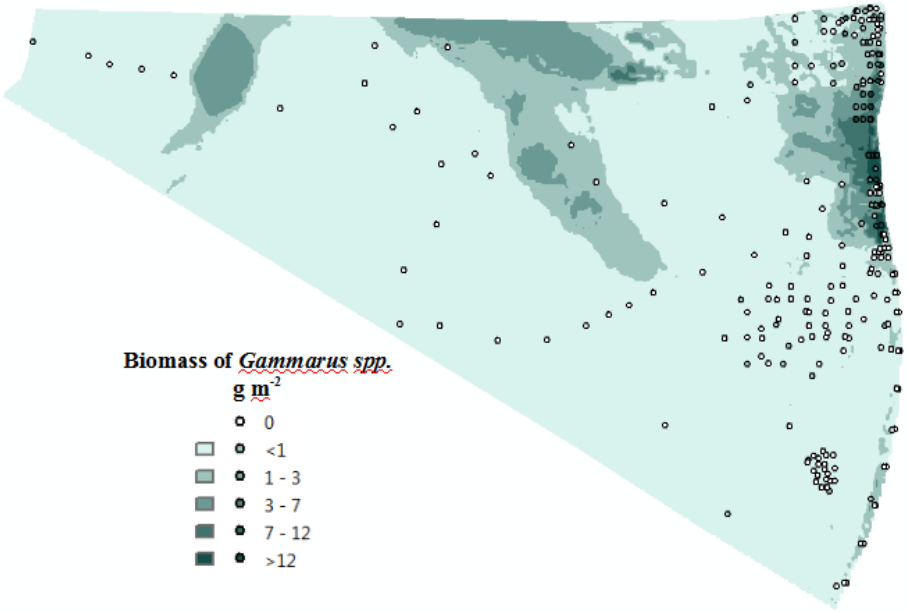

Figure 6. Sampled and modelled biomass of Gammarus spp.

However, hard bottom was sampled only in the photic coastal area, so other two high biomass areas in deeper parts of LEZ were not ground-proved and should be interpreted with appropriate caution. There are rare occasions of gammarid occurrence in soft-bottoms. This may be because of grab sampling is not the best technique for sampling of these amphipods. It seams that they can escape the grab just before it is closed.

2.2.2 Halicryptus spinulosus

Priapulid *H.spinulosus* occurs almost in all offshore area with fine sediments (Figure 7). Low biomass areas are sandy sediments at 20-40 meter depth, while highest biomass are found in muddy sediments at 50-60 meter depth.

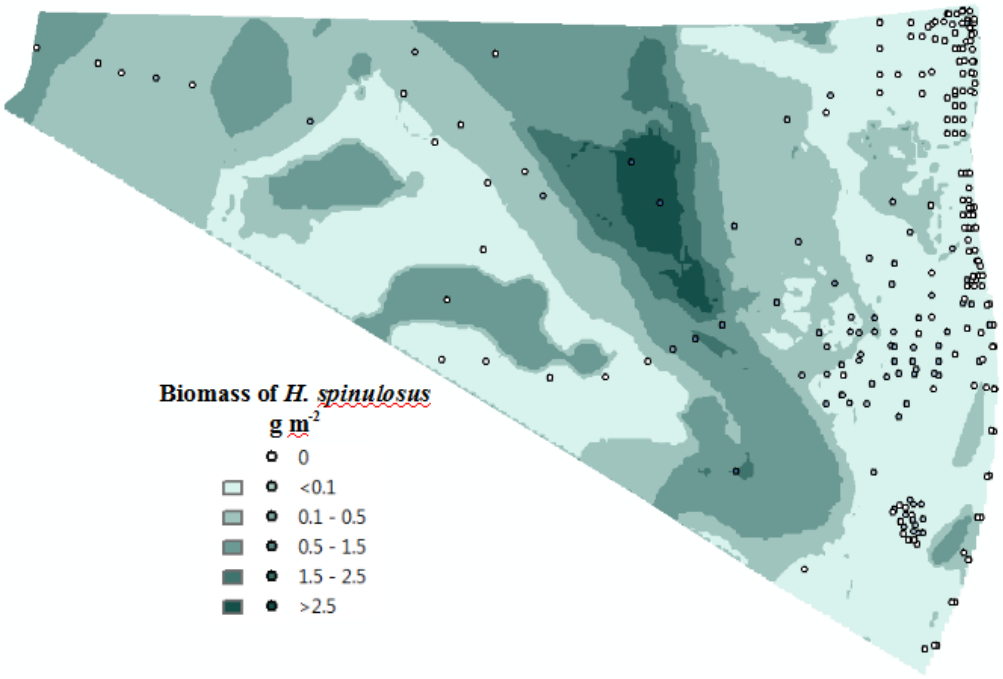

Figure 7. Sampled and modelled biomass of Halicryptus spinulosus

The same as for other benthic macrofauna hypoxia sets the depth limit for *H.spinulosus*, however that is not reflected in the model due to the low sampling effort in the deepest part of LEZ. The modelling task was aggravated by the specificity of sample biomass. In most cases of *H.spinulosus* presence in samples, biomass values were relatively low: in 94 % of sampling cases priapulids weighted less than 1 g m^{-2}. In other 6 % cases biomass values were up to 8 times higher. This uneven biomass distribution gave the unnecessary bias for the model thus increasing its mean absolute error.

2.2.3 Hediste diversicolor

The highest biomass of *H.diversicolor* is found in coastal area in both soft and hard bottoms (Figure 8). Relatively high abundance of this polychaete can be found down to 50 meters, but there are rare cases, where it was found down to 70 meters or even at 120 meter depth. Probably due to the patchy distribution the model we not as accurate as for other zoobenthos species (the lowest accuracy).

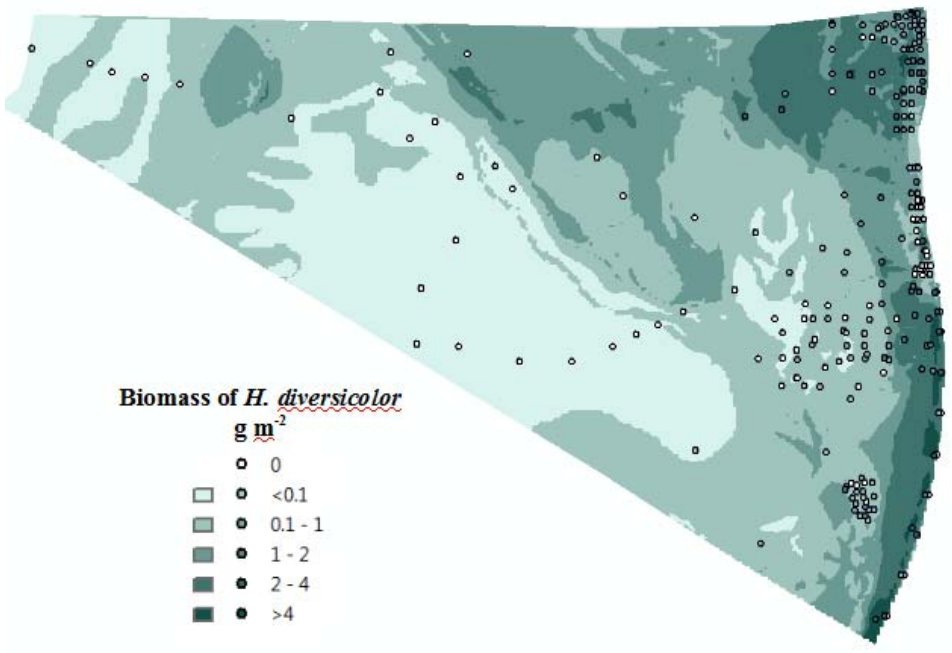

Figure 8. Sampled and modelled biomass of Hediste diversicolor

2.2.4 Macoma balthica

Bivalve mollusc *Macoma balthica* is one of the most common benthic invertebrate in LEZ. While juvenile molluscs can occur in the very shallow areas in almost all types of sediments, adults occupy nearly all soft-bottom areas roughly from 10 meters down to hypoxic conditions (Figure 9).

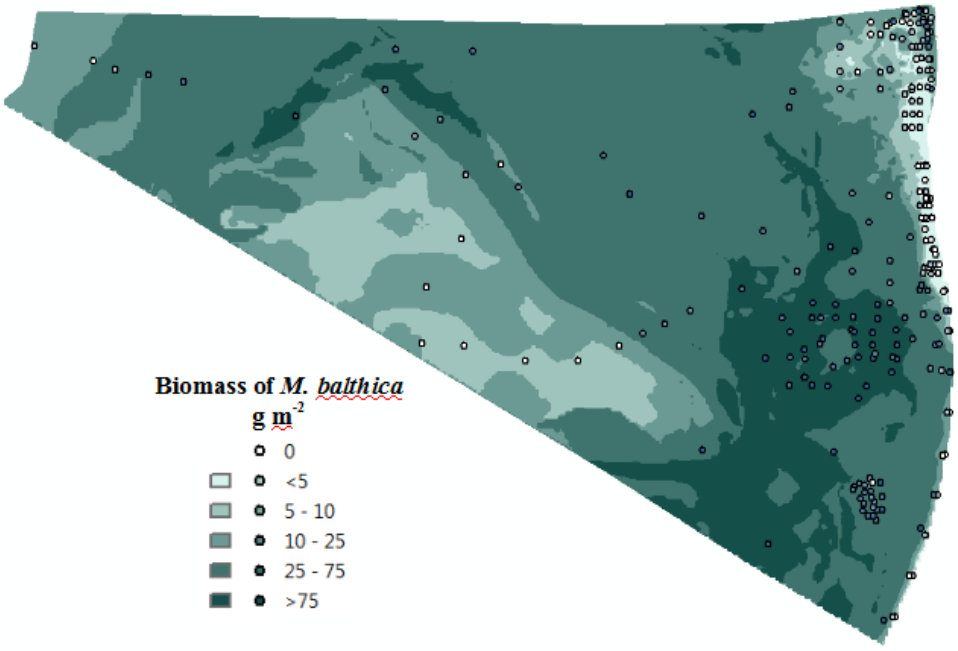

Figure 9. Sampled and modelled biomass of Macoma balthica

It is fair to notice that in the area where *M.balthica* is absent (South-central part of LEZ) model predicts biomass up to 10 g m^{-2}. This is because the range (0-253 g m^{-2}) is many times

bigger that the mean absolute error (26.4 g m^{-2}). So to interpretation of the map of biomass distribution of *M.balthica* for presence-absence borders or low biomass areas must be with appropriate caution, however higher biomass areas are delineated adequately.

2.2.5 Marenzelleria neglecta

Similar to the native polychaete *H.diversicolor* the highest biomass of invasive *M.neglecta* were predicted in the coastal area also on both soft and hard substrate (Figure 10). This polychaete worms are found down to 50-60 meters, rarely up to 80 meter depth. Probably due to the patchy distribution the model we not as accurate as for other zoobenthos species (second lowest accuracy).

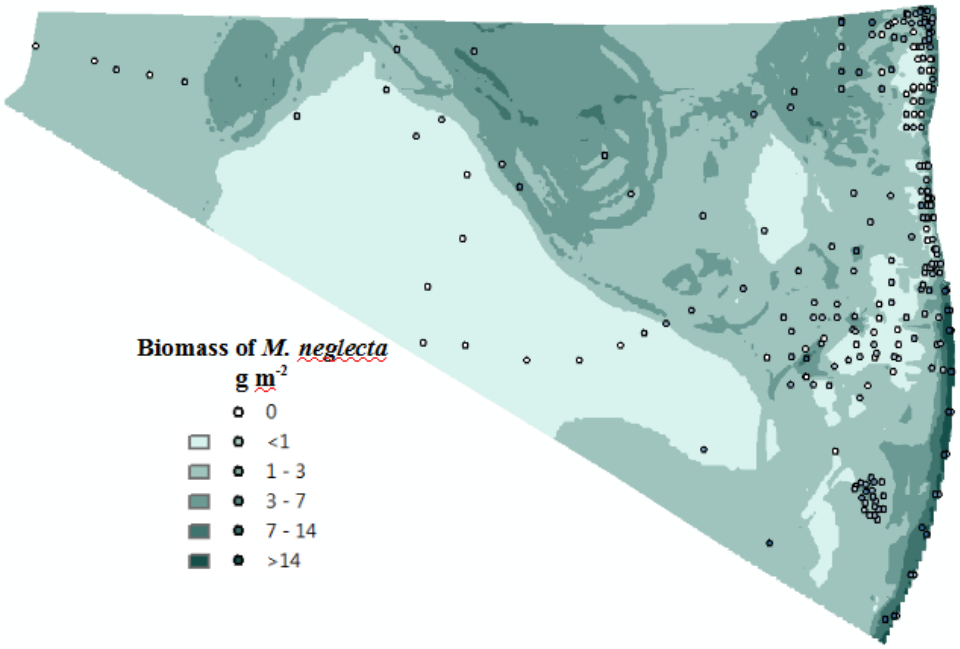

Figure 10. Sampled and modelled biomass of Marenzelleria neglecta

2.2.6 Mya arenaria

Adults of this mollusc species mostly occur roughly from 10 down to 35 meters, rarely down to 45-50 meter depth (Яарвекюльг, 1979, Olenin, 1997). Juveniles, however, are found mostly in coastal area, can dwell in soft sediments or be attached to macroalgae of stones. Unlike *M.balthica*, juveniles of *M.arenaria* can occur in relatively high abundance and biomass. The overall output of model (Figure 11) can be trusted but only down to 50 meter isobath which is a natural border of the occurrence of this species. We can tolerate small biomass values in the South-central part for the same reasons as in *M.balthica* case, however moderate biomass predictions in the deep most Western part of LEZ must be treated as errors. This happened probably because the depth as a environmental predictor was excluded due to cross-correlations with other predictors (explained in paragraph 1.2.3).

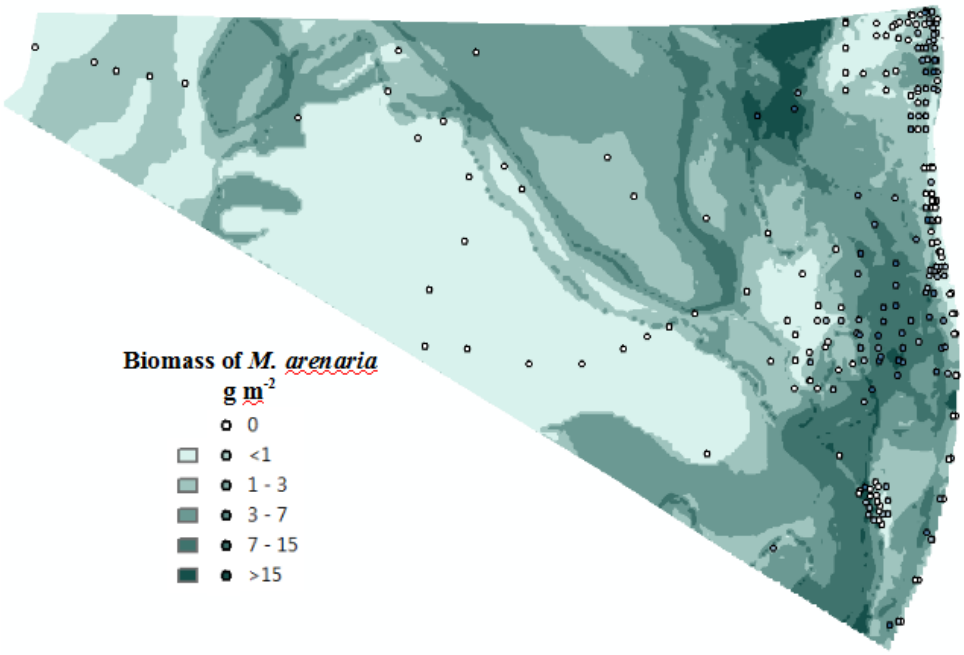

Figure 11.Sampled and modelled biomass of Mya arenaria

2.2.7 Mytilus edulis

As expected highest biomass of *M.edulis* were predicted in the stony bottom of the coastal zone. As mentioned in paragraph 2.1.2 the model of blue mussel had the best accuracy even if areas of biomass less than 100 g m^{-2} (Figure 12) should be interpreted as absence areas or exceptional cases due to the same reasons as in *M.balthica* case described in paragraph 2.2.4.

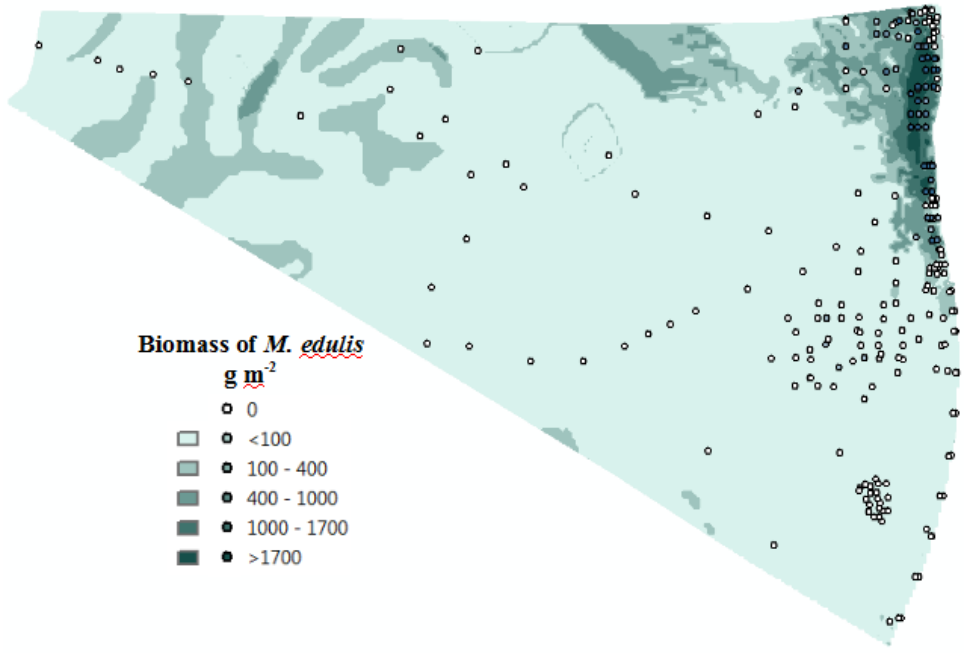

Figure 12. Sampled and modelled biomass of Mytilus edulis

The biggest drawback of this model is that biomass distribution of *M.edulis* is built-up only according to field data from coastal zone (in depth range of 0-20 meters). So deeper hard-bottom areas are not ground-proved thus interpretation should be with appropriate caution.

2.2.8 Saduria entomon

The output map of Modelled biomass of *S.entomon* clearly show that this isopod are most abundant in the soft-bottom offshore area (Figure 13).

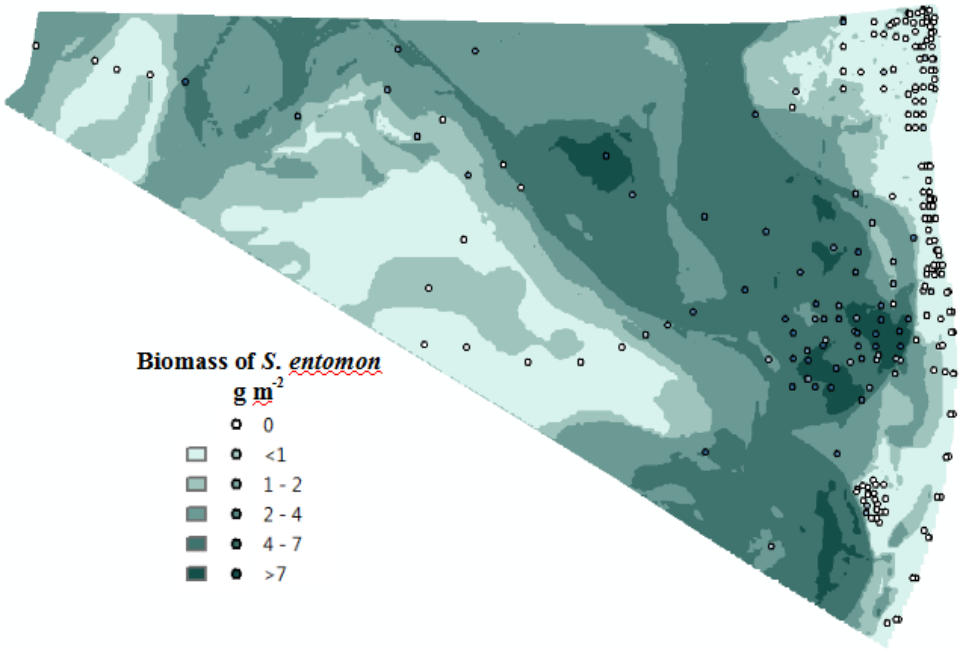

Figure 13. Sampled and modelled biomass of Saduria entomon

The depth limit for the field data was 67 meters and the model recorded this border quite well, however an artefact occurred in depths of 110-120 meters (the westernmost part of the LEZ) where the model predicted moderate biomass values (2-4 g m^{-2}). Nonetheless, the overall model output is reasonable and could be interpreted without many exceptions.

2.3 Zonation of fish feeding grounds

As described in paragraph 1.3.1 different weights are assigned according to *occurrence* and *importance* of prey items (Figure 14) for every fish species separately.

Prey items	Cod	Flounder	Eelpout
Gammarus spp.	1	1	0.75
Halicryptus spinulosus		0.5	
Hediste diversicolor	0.37	0.37	0.37
Macoma balthica		1	0.5
Marenzelleria neglecta	0.25	0.25	0.25
Mya arenaria		0.25	
Mytilus edulis		0.5	0.25
Saduria entomon	1	0.37	0.37

Figure 14. Weight multipliers of prey items assigned according to occurrence and importance. Free cells indicate that fish do not prey upon particular item

Judging from Figure 14 it is easy to sort prey items by their importance for the feeding of fish. Baltic cod mainly prey upon gammarids and isopods *S.entomon*, while polychaete worms are of minor importance. Preferred prey items for flounder and eelpout were gammarids and bivalves *M.balthica*, while priapulids *H.spinulosus* and soft-shell clams *M.arenaria* were eaten only by flounder. Flounder had the most diverse diet composition (total of eight prey items), while eelpout and cod preyed upon six and four prey items respectively. Finally, maps of seabed importance for feeding of cod, flounder and eelpout (Figures 15-17) are derived by multiplying biomass maps of prey items by assigned weights and adding them up.

The important areas of the seabed for the feeding of Baltic cod (Figure 15) are distributed down to 50 meter depth, while 60 meter isobath almost perfectly separates important areas from unimportant. One of the most important areas in the Northern part of the coastal zone are determined by hard substrate seafloor where several species of gammarids are present in high abundance and biomass. Other two very important areas are situated in the central and westernmost part of LEZ. These two are mostly determined by high biomass of glacial relict *S.entomon*.

The most important area for the feeding of flounder are situated in the Northern part of the coastal zone (Figure 16). This area is determined by incomparably high biomass of hard-substrate associated blue mussel *Mytilus edulis*, even if this mollusc is moderately important for the feeding of flounder. However flounder is present (thus also feeding) down to 80 meters or more, so in order to delineate important areas in the offshore zone *M.edulis* were excluded from the map (small map in Figure 16). This map shows that the most important soft bottom areas are distributed from 20 down to 60 meter depth. Shallower areas are mostly determined by infaunal clams and polychaetes, while deeper ones by *M.balthica* and *S.entomon*.

The most important areas for feeding of eelpout (Figure 17) are very similar to the flounder case, probably because of similar diet composition. However, eelpout is a coastal fish so the focus must be on near-shore zone before delineating areas of importance. It is obvious that most important areas for eelpout are coastal hard substrate area from 5 down to 20 meter depth situated in the North-eastern part of LEZ. It is hard to delineate important areas in soft-bottom seafloor (small map in Figure 17) due to more or less homogeneous distribution of prey items.

The map of overall importance of the seabed for all three fish species is presented in Figure 18. The most important area for all three fish species is the stony bottom in the coastal zone situated in the Northernmost part of LEZ. There are some important areas in the offshore zone for cod and flounder situated in the central and Western parts of LEZ.

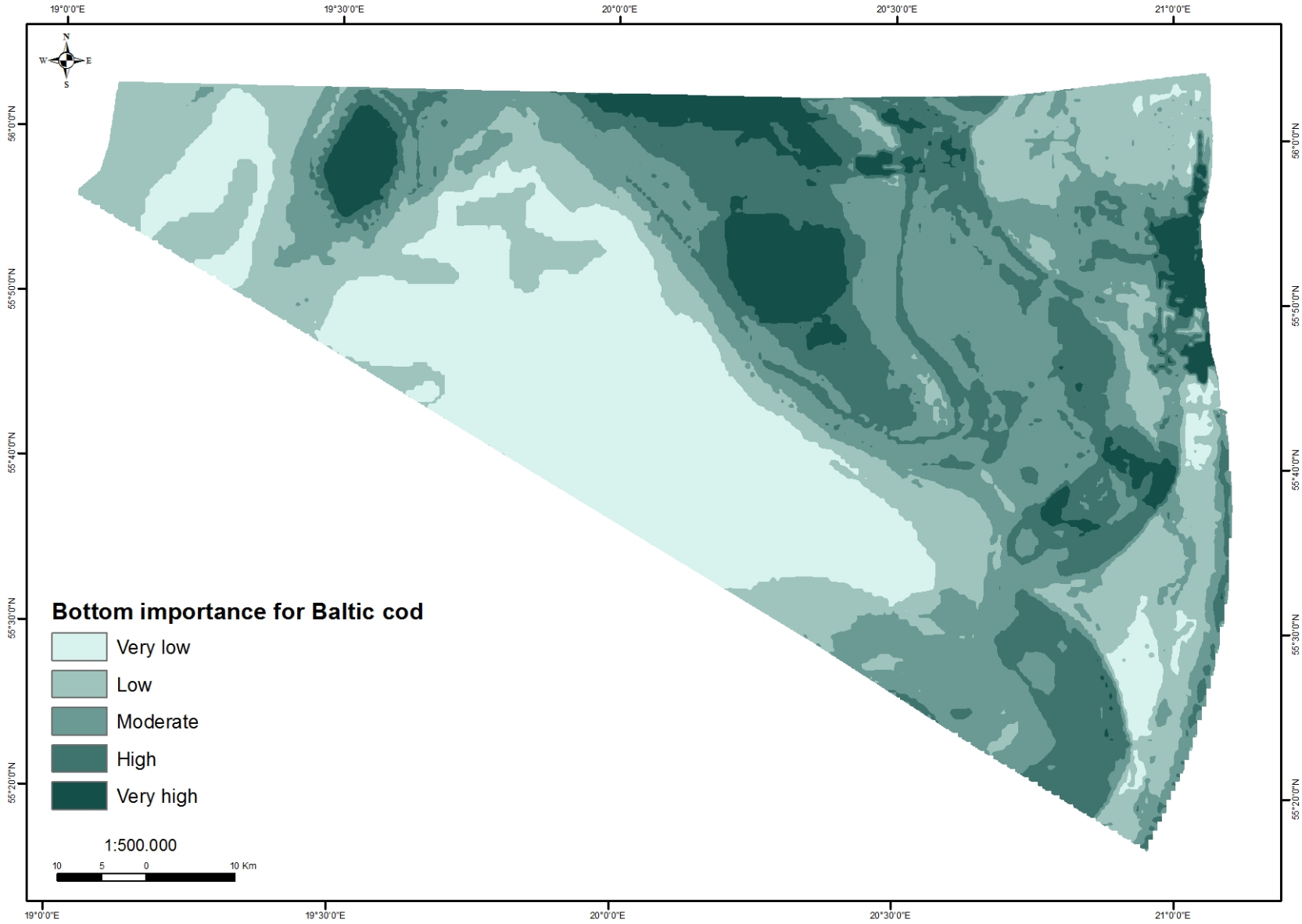

Figure 15. Seabed quality for feeding of Baltic cod based on the biomass distribution of prey items

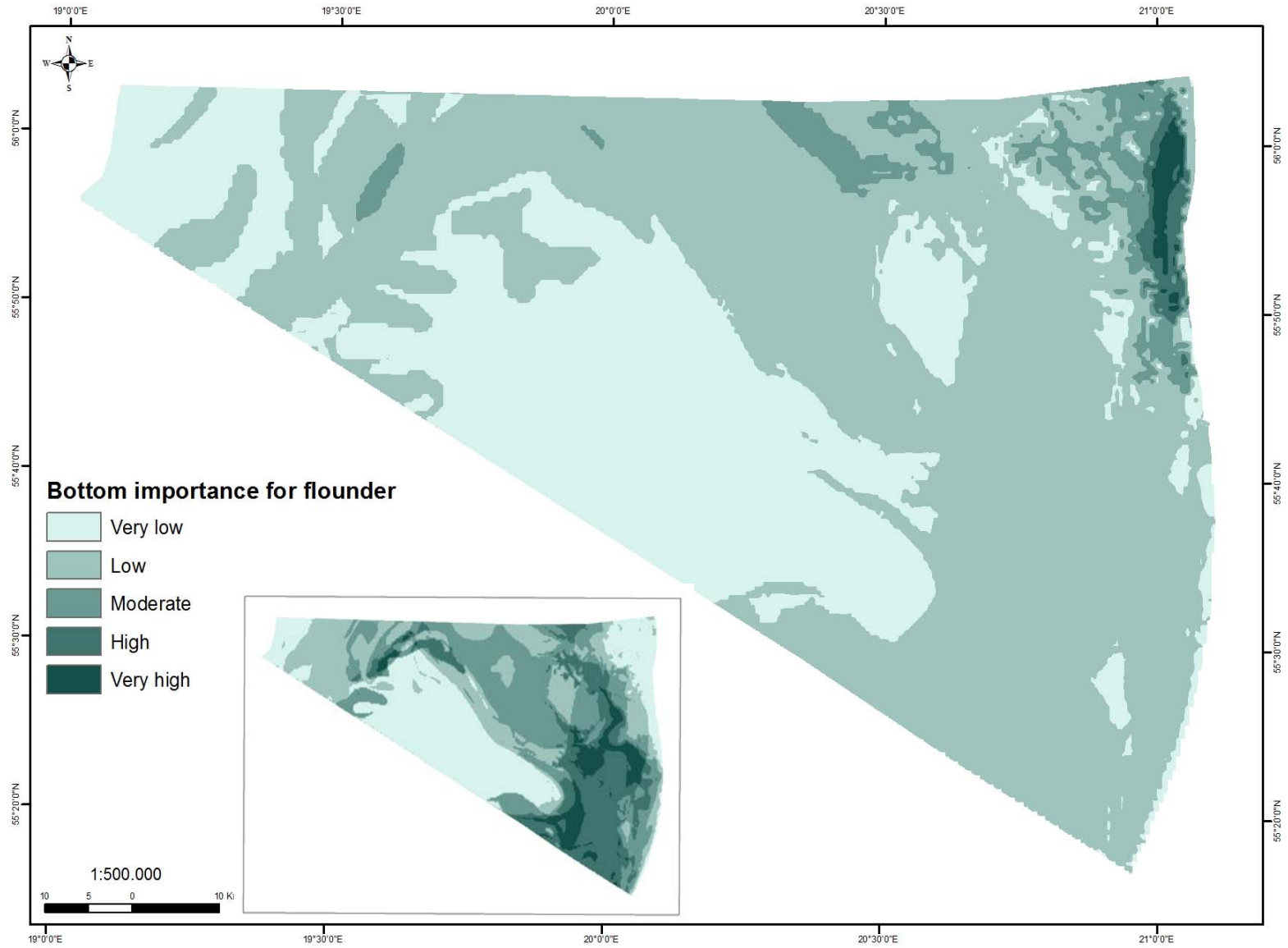

Figure 16. Seabed quality for feeding of flounder based on the biomass distribution of prey items. Smaller map shows the same importance excluding Mytilus edulis

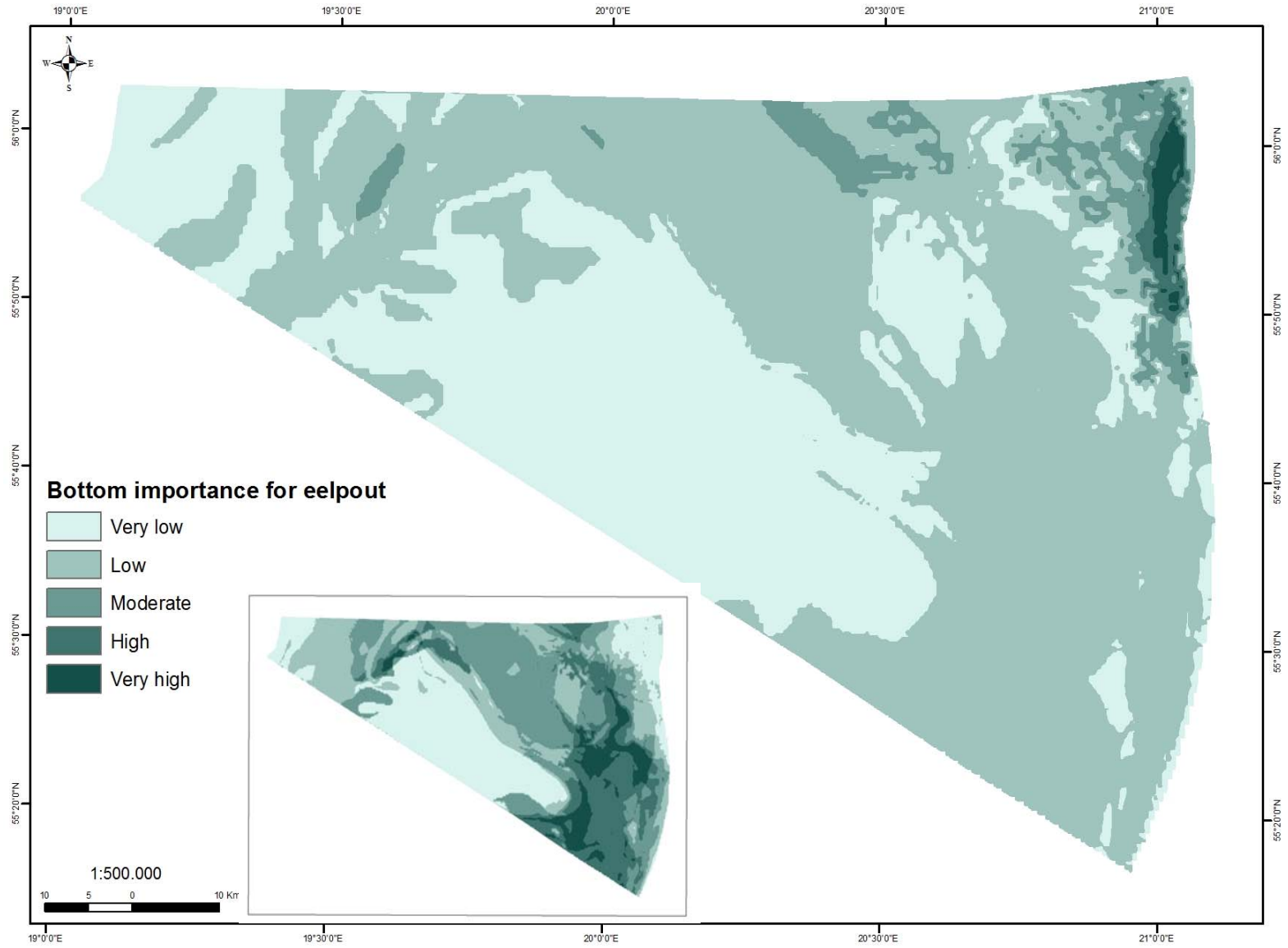

Figure 17. Seabed quality for feeding of eelpout based on the biomass distribution of prey items. Smaller map shows the same importance excluding Mytilus edulis

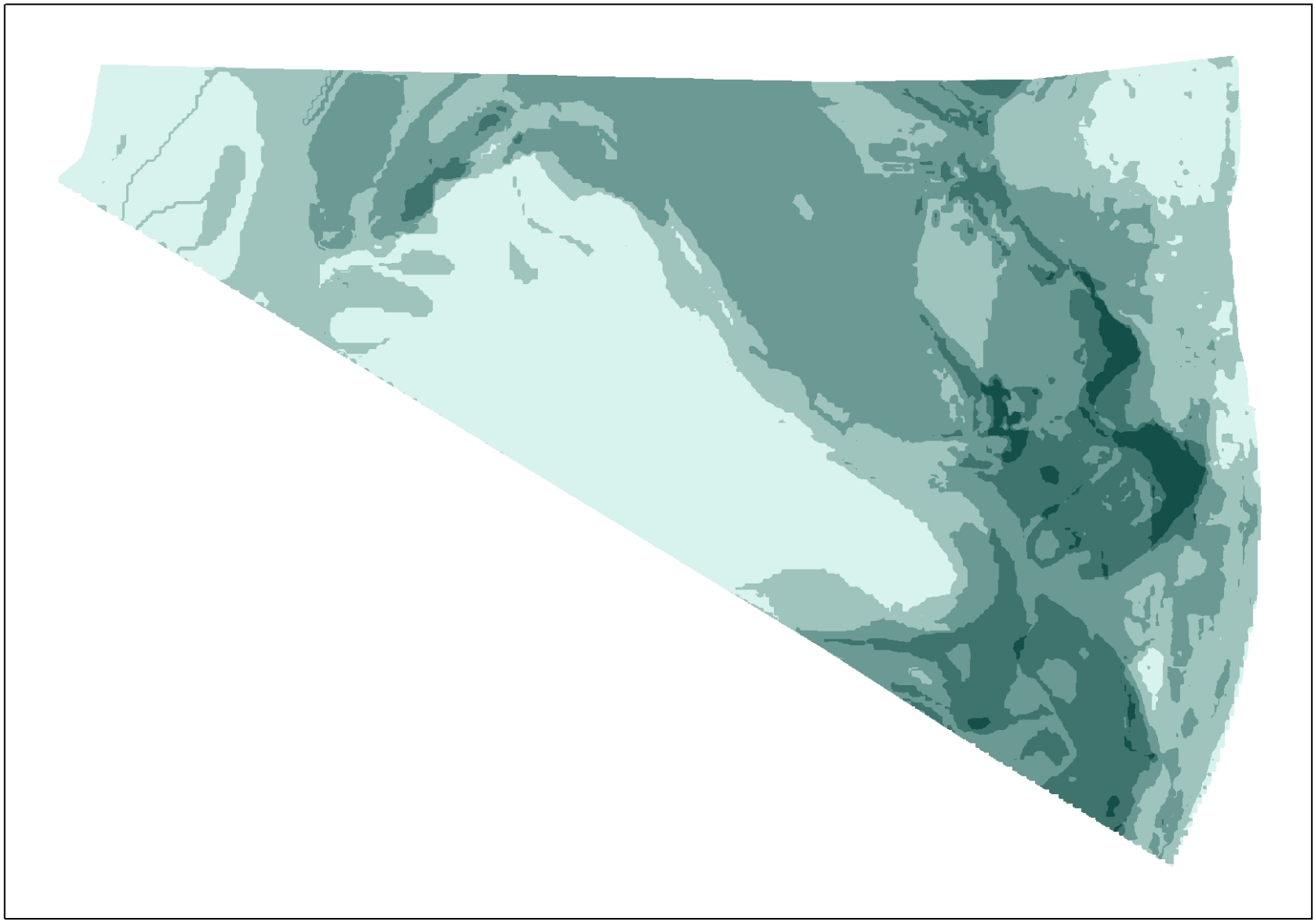

Figure 18. Seabed quality for feeding of cod, flounder and eelpout based on the biomass distribution of prey items

2.4 Accuracy assessment of prediction maps

As described in paragraph 1.3.2 the accuracy assessment (Figure 19) is based on different number of field samples in separate intervals or categories of environmental predictors and uneven importance of these predictors for modelled prey items.

"*High*" accuracy should be interpreted as the best possible modelled area with a current dataset, though validation errors still must be included. Areas of *"moderate*" accuracy should be treated as trustworthy, however they should be double-checked or ground-proved before decision making. "*Low*" accuracy indicates areas that are modelled based on few samples and must be treated accordingly.

On the other hand, accuracy levels can be useful for planning future data sampling strategy. In that context "*high*" accuracy would mean sufficient data (if validation errors are reasonable) and "low" accuracy would indicate areas that should be sampled the most.

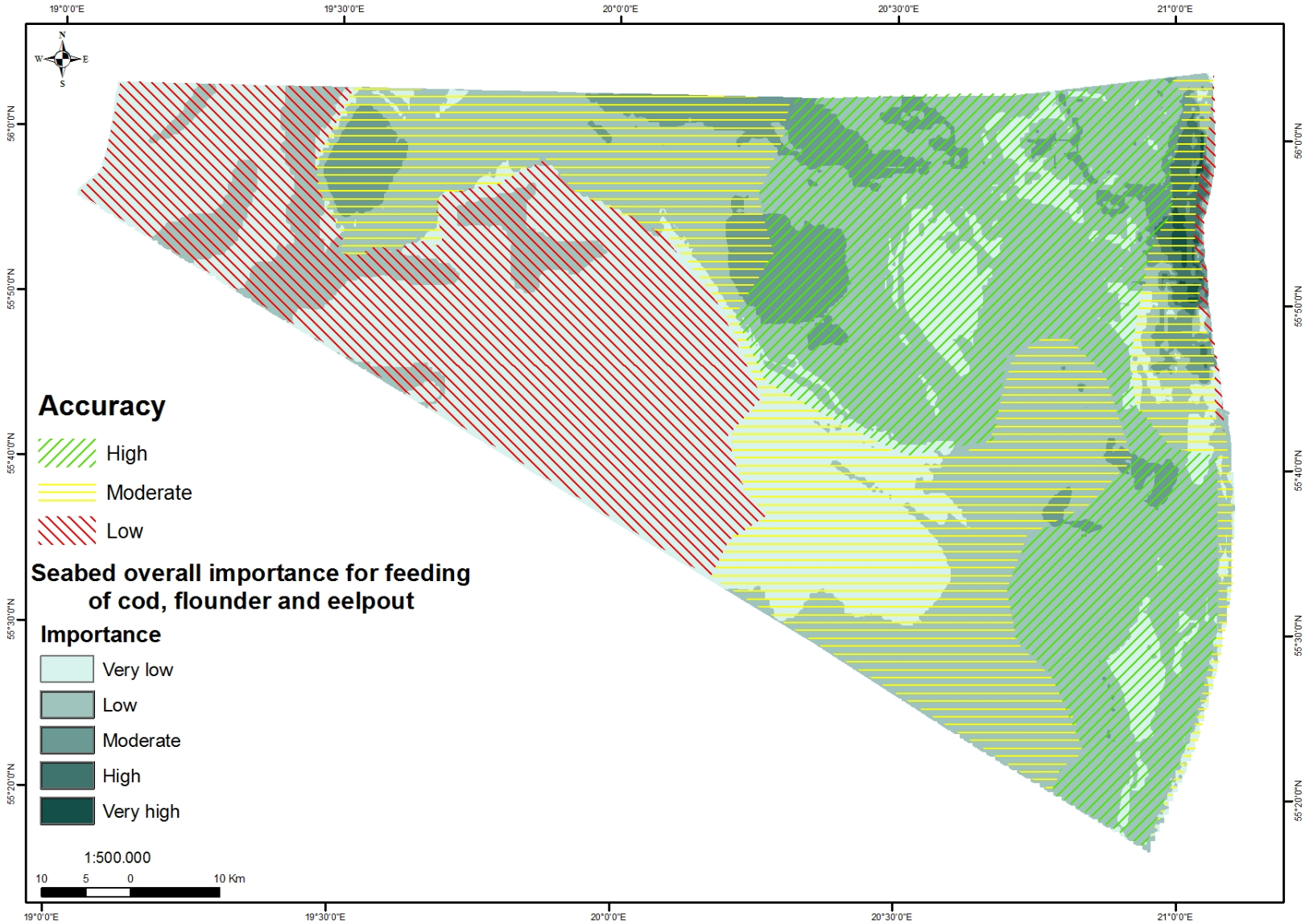

Figure 19. Accuracy levels for the map of seabed overall importance for feeding of cod, flounder and eelpout based on the biomass distribution of prey items.

REFERENCES

Breiman L. 2001. Random Forests. Mach Learn 45:5–32.

Friedman, J. (2001). Greedy function approximation: the gradient boosting machine, *Ann. of Stat.*

HELCOM,1988. Guidelines for the Baltic monitoring programme for the third stage. Part D. Biological determinants, 23-87.

Kuhn,S., Egert,B., Neumann,S., Steinbeck.,Ch. 2008. Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. BMC Bioinformatics 2008, 9:400

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2, 18–22.

Olenin, S. 1997. Benthic zonation of the Eastern Gotland Basin. Netherlands Journal of Aquatic Ecology, Vol. 30, No. 4: 265-282.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9, 181–199.

Repečka, R. 2003. The species composition of the ichthyofauna in the Lithuanian economic zone of the Baltic Sea and the Curonian lagoon and its changes in recent years. Acta Zoologica Lituanica, 2003, Volumen 13, Numerus 2

Vincenzi S., Zucchetta M., Franzoi P., Pellizzato M., De Leo G.A., Torricelli P. 2011. Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. Ecological Modelling 222 (2011) 1471–1478

Яарвекюльг, А., 1979. Донная фауна восточной части Валтийскогоморя. Таллин "Валгус", - Р.324.