since 1961

Baltica
http://www.geo.lt/geo/index.php?id=71C

BALTICA Volume 25 Number 2 December 2012 : 163–170
doi:10.5200/baltica.2012.25.16

Species distribution modelling of benthic invertebrates in the south-eastern Baltic Sea

Andrius Šiaulys, Martynas Bučas

Šiaulys, A., Bučas, M., 2012. Species distribution modelling of benthic invertebrates in the south-eastern Baltic Sea. *Baltica, 25 (2), 163-170.* Vilnius. ISSN 0067-3064.

Manuscript submitted 17 September 2012 / Accepted 19 November 2012 / Published online 10 December 2012

Abstract The distribution of benthic invertebrates is one of the key parameters for the marine spatial planning and management, however traditionally the data on benthic invertebrates are based on point sampling. Recently statistical methods of predictive modelling are used to create maps of species distribution, nevertheless, no comparative analysis of different modelling methods has been yet performed in the Baltic Sea region. In this study the occurrence and biomass distribution of 23 benthic species in the southeastern Baltic Sea were modelled. A comparison of the following predictive modelling methods was performed: random forests (RF), generalized additive models (GAM), multivariate adaptive regression splines (MARS) and maximum entropy (MaxEnt). In order to assess the consistency of the methods, 100 iterations with different train/test datasets were made for each of them. Random forests achieved the highest predictive performance for both species occurrence and biomass distribution models; also it was the most consistent for different iterations. Predictive performance of GAMs and MARS followed RF, whereas MaxEnt accurately predicted occurrence only for the species with a relatively low distribution range.

Keywords *• Marine benthic invertebrates • Random forests • Generalized additive models • Multivariate adaptive regression splines • Maximum entropy • Lithuanian waters area • south-eastern Baltic Sea*

✉ *Andrius Šiaulys [andrius@corpi.ku.lt], Martynas Bučas, Coastal Research and Planning Institute, Klaipėda University, Herkaus Manto g. 84, LT-92294, Klaipėda, Lithuania.*

INTRODUCTION

Marine spatial planning and management require spatial information on environmental characteristics (Foley *et al.* 2010; Allnutt *et al.* 2011, Guerry *et al.* 2012). However, they are usually based on the point data, especially for the distribution of marine biota. Sampling sites rarely are dense and evenly distributed within study area to use simple interpolation techniques for the creation of spatial maps (Li, Heap 2008). Species distribution models (SDMs) relate the occurrence or abundance of organisms with the environment factors that limit their distribution and can predict the species potential habitat using environmental data. SDMs gained an increasing attention in recent years followed by many applications in aquatic ecology (Robinson *et al.* 2011) from global predictions of the seafloor biomass (Wei *et al.* 2010) to species distributions at regional scale (Gogina, Zettler 2010; Vincenzi *et al.* 2011) or even mapping ecosystem services (Šiaulys *et al.* 2012).

Various modelling techniques are used for modelling species distribution (Guisan, Zimmerman 2000; Elith *et al.* 2006). However, only few SDMs techniques have been applied in the Baltic Sea, especially, modelling the distribution of benthic invertebrates (Carlström *et al.* 2010; Gogina, Zettler 2010; Šiaulys *et al.* 2012). The studies used different SDMs methods and no comparison of methods has been performed. Therefore, in this study a comparison of four modelling techniques for the prediction of 23 benthic invertebrate species in the south-eastern part of the Baltic Sea has been performed: generalized additive models (GAM), multivariate adaptive regression splines (MARS), maximum entropy (MaxEnt) and random forests (RF). Moreover, the variance of model performance was estimated by iterating random data splitting into train and test datasets. Different SDMs were proposed by

Kuhn *et al.* (2008) and Reiss *et al.* (2011), however, the most suited model can depend on the data traits (Elith 2011), such as species distribution range or prevalence of species occurrence (Manel *et al.* 2001). The later effects have been tested in this study.

MATERIAL AND METHODS

Environmental predictors and field data

This study was carried out in the Lithuanian Exclusive Economic Zone, southeastern Baltic Sea (Fig. 1). Of the available environmental predictors known to be important for the distribution of benthic invertebrates (Olenin 1997; Bučas *et al.* 2009; Gogina, Zettler 2010; Reiss *et al.* 2011), eight were used for the models of species occurrence and biomass distribution: sediment types, Secchi depth, minimum near-bottom oxygen concentration, near-bottom current velocity, wave generated orbital near-bottom velocity, slope and roughness of the seabed, areas of above and below the thermocline. Quantitative environmental parameters were tested for collinearity and predictors were removed from models if the variance inflation factors were > 3 (Quinn, Keough 2002). The layers of sediments, slope and roughness were derived from geological and bathymetrical charts ([1]Repečka *et al.* 1997; Gelumbauskaitė *et al.* 1999; Bitinas *et al.* 2004). Sediments were classified into four types: boulders, cobbles/gravel, sand and silt (Wentworth 1922). The wind wave orbital velocity data layer was derived using SWAN model (Booij *et al.* 1999) based on 2008–2009 wind data. National marine monitoring data was used to derive Secchi depth and thermocline layers (MRC, unpublished 1998–2006). The mean annual minimum near-bottom oxygen concentration (2000-2006) and bottom current velocity layers were derived from datasets produced by BALANCE project (Hansen *et al.* 2007; Bendtsen *et al.* 2007).

The dataset of the study consists of 640 benthic samples taken at 224 sampling sites during 1998–2010 (Fig. 1). Soft-bottom samples were taken with a Van-Veen grab, while hard bottoms were sampled by SCUBA divers with 0.20 x 0.20 m frame. These samples were taken and treated following standard guidelines for bottom invertebrate sampling (HELCOM 1988).

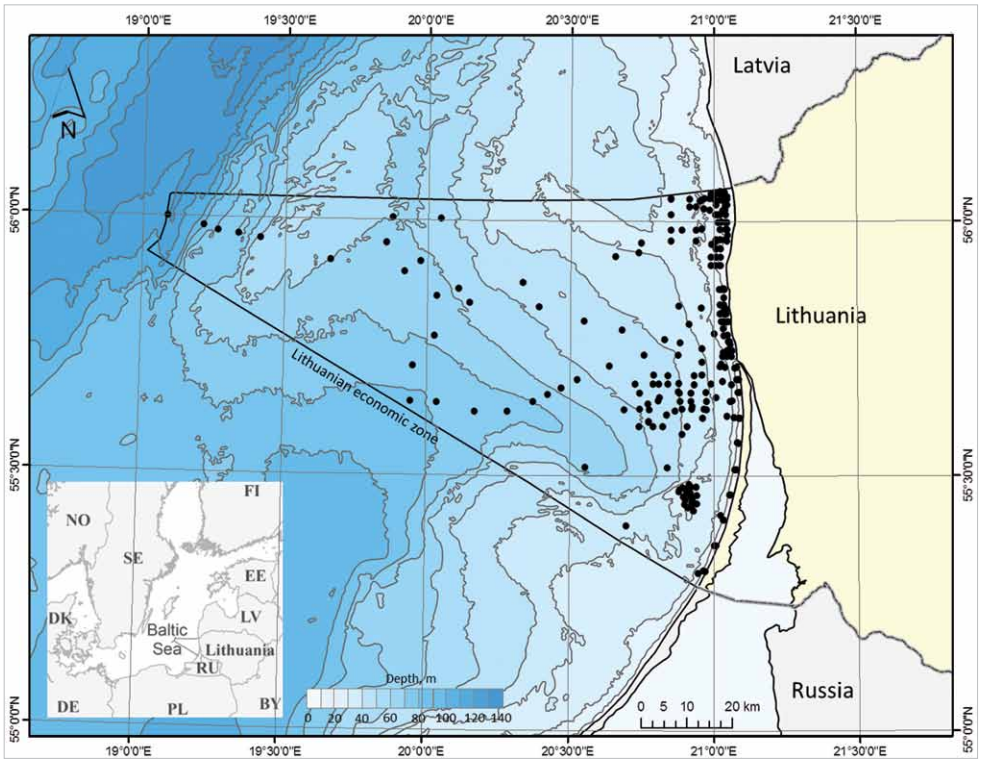

Fig. 1 Location of the sampling sites in Lithuanian waters in 1998–2010. Compiled by A. Šiaulys, 2012 (bathymetry acquired after L. Ž. Gelumbauskaitė 2009).

Modelling techniques

Generalized additive models (GAM)

GAMs are semi-parametric extensions of generalized linear models with the assumption that the functions are additive and that the components are smooth. This method deals well with the highly non-linear and non-monotonic relationships between the set of explanatory and response variables (Guisan *et al.* 2002). Model selection was based on penalized regression splines with default gamma-values and a maximum four degrees of freedom for continuous predictor variables in order to maintain ecologically interpretable models (Wood, Augustin 2002). The "mgcv" 1.7-9 package (Wood 2006) within R environment was used for occurrence and biomass distribution models.

Multivariate adaptive regression splines (MARS)

MARS algorithm is a nonparametric method for multiple regression, which uses adaptively selected spline functions (Hansen, Kooperberg 2002) developed by Friedman (1991). MARS is based on linear relationships, however it identifies and estimates a model which coefficients differ depending on the level of the predictor variable (Reiss *et al.* 2011). Models were built using the GLM approach and specified to

1 Repečka, M., Gelumbauskaitė, Ž., Grigelis, A., Šimkevičius, P., Radzevičius, R., Monkevičius, A., Bubinas, A., Kasperovičienė, J., Gadeikis, S., 1997. National marine geological mapping at a scale of 1:50 000, Klaipėda–Šventoji water area, Object No. I. Manuscript, Lietuvos geologijos tarnyba, Lietuvos geologijos institutas, Vilnius, 227 pp. [In Lithuanian].

include first order interactions, where significant. Both occurrence and biomass distribution models were built using the "earth" package (Milborrow 2012) under R environment.

Maximum entropy (MaxEnt)

MaxEnt is a general-purpose machine learning method which estimates a target probability distribution by finding the probability distribution of maximum entropy and constraining the expected value of each environmental variable to match its empirical average (Phillips *et al.* 2006; Reiss *et al.* 2011). In this study we used MaxEnt program version 3.3.3e (Philips *et al.* 2006; Philips *et al.* 2008). The convergence threshold was set at 10^{-5} and the maximum number of iterations at 500 to allow the algorithm to get close to convergence (Phillips *et al.* 2006). Although, MaxEnt works well with presence-only datasets (Elith *et al.* 2011), absence data was also used in MaxEnt models to be more consistent with other methods. MaxEnt was used only for modelling the occurrence probability of species.

Random Forests (RF)

RF is a classification and regression model developed by Breiman (2001) that generates multiple classification trees with a randomised subset of predictors (Reiss *et al.* 2011). A large number of trees are grown and the number of predictors used to find the best split at each node is a randomly chosen subset of the total number of predictors (Prasad *et al.* 2006). In this study the number of trees was set to 1000, the number of variables randomly selected at each node and minimum node size were set to default values. The "randomForest 4.6-2" package (Liaw, Wiener 2002) within the R environment was used for predictions of presence probability and biomass distribution of benthic species.

Predictive performance, model variation and effects of data traits

Predictive performance of the species occurrence models was estimated by area under the receiver operating characteristic curve (AUC) measures. The AUC values range between 0 and 1. According Hosmer and Lemeshow (2000) "excellent" prediction performance is achieved when AUC > 0.9, "good" performance – AUC 0.7-0.9, "poor" performance – AUC < 0.7. If AUC is ≤ 0.5 then predictions are no better than random. For biomass distribution models two measures were estimated: root mean square error normalized by range (NRMSE) and coefficient of determination (R^2).

The initial dataset was split into train set used for model build-up (70% of data) and test set used for validation (rest 30% of data) ensuring that species prevalence (the ratio between sites where a particular species is present and total number of sites) would be in equal proportions in train and test datasets. The variance of model performance was assessed by mean values of 100 iterations of splits for each species. Variation is expressed by the coefficient of variation CV_{AUC} for occurrence models and CV_{NRMSE} and CV_{R}^{2} for biomass distribution models.

The effect of data traits, species prevalence and distribution range, on both predictive performance and model variation were tested using Pearson's correlation between AUC, NRMSE, R^2. The species distribution range was determined for each species using convex hull algorithm in Quantum GIS 1.7.4 (Quantum GIS Development Team, 2010). The species prevalence and distribution range significantly correlated ($r = 0.70$, $p < 0.01$), therefore the effect of the species prevalence was tested on the predictive performance of models.

RESULTS

Performance of models

All four methods on average achieved "good" predictive performance for occurrence models (Fig. 2). The highest performance was achieved by RF (AUC = 0.87 ± 0.06), followed by GAM (AUC = 0.84 ± 0.06), MARS (AUC = 0.80 ± 0.06) and MaxEnt (AUC = 0.77 ± 0.11). RF models were also the most consistent ranging from "good" to "excellent" performance (AUC = 0.78–0.96), closely followed by GAM (AUC = 0.74–0.95) and MARS (AUC = 0.70–0.95), while MaxEnt (AUC = 0.56–0.93) had six cases of "poor" predictive performance. According to coefficients of variation of AUC the most consistent method was again RF (CV_{AUC} = 0.05±0.02), closely followed by MaxEnt (CV_{AUC} = 0.06±0.02) and GAM (CV_{AUC} = 0.06±0.03), while MARS (CV_{AUC} = 0.09±0.04) varied the most.

Fig. 3 indicates that the mean prediction error of all three methods for the biomass distribution was very similar (NRMSE = 0.08± 0.04) among the methods. According to the coefficient of determination (R^2) the best mean performance was achieved by RF (R^2 = 0.32 ± 0.19), followed by MARS (R^2 = 0.13 ± 0.14) and GAM (R^2 = 0.12 ± 0.12).

As shown in Fig. 4 the R^2 of biomass distribution models by RF was the most consistent (CV_{R}^{2} = 0.45±0.29), but with the highest variance of NRMSE (CV_{NRMSE} = 0.80±0.39). GAMs were relatively consistent according to R^2 (CV_{R}^{2} = 0.88±0.41) and NRMSE (CV_{NRMSE} = 0.52±0.25), while the models of MARS were consistent in respect of NRMSE (CV_{NRMSE} = 0.51±0.23), but with the high variance in R^2 (CV_{R}^{2} = 1.30±0.77).According to AUC, all methods correlated with each other ($r = 0.66$-0.90, $p < 0.01$), except for a weak correlation between MARS and MaxEnt (Table 1). Very strong correlations were estimated among

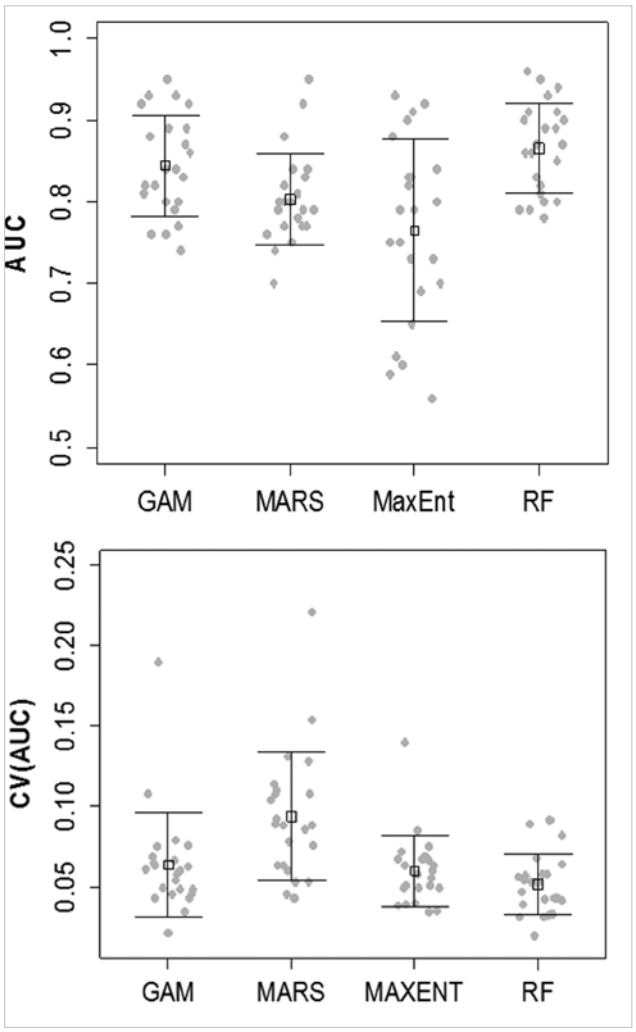

Fig. 2 Performance of four predictive methods for modelling of the occurrence of benthic invertebrates according AUC (area under the curve) values and coefficients of variation of AUC during 100 of iterations. Compiled by A. Šiaulys and M. Bučas, 2012.

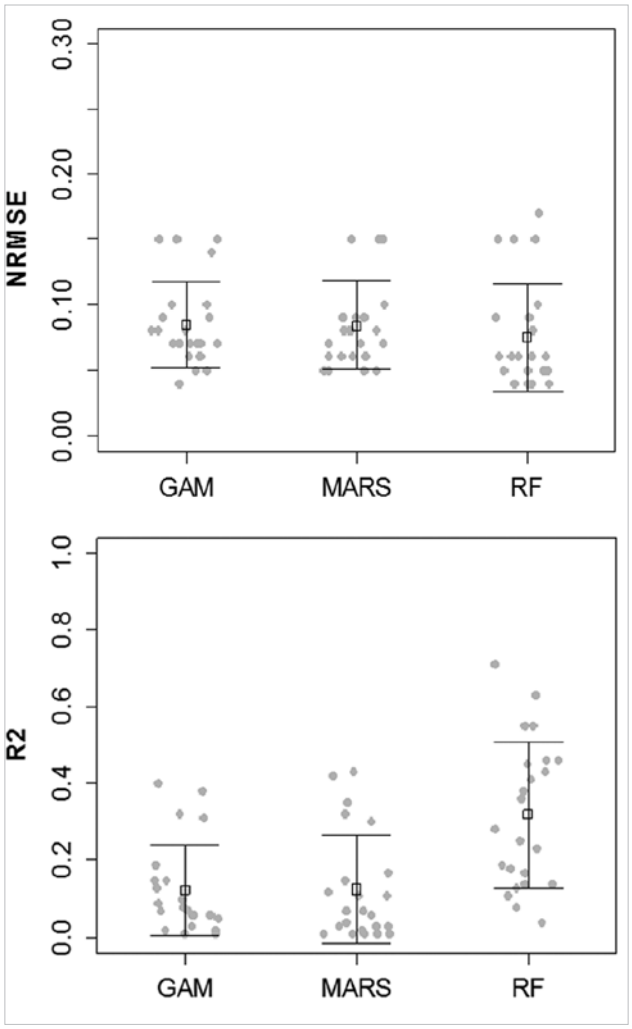

Fig. 3 Performance of three predictive methods for modelling of the distribution of benthic invertebrates biomass according NRMSE (root mean square error normalized by range) and R^2 (coefficient of determination) values. Compiled by A. Šiaulys and M. Bučas, 2012.

the methods in respect of NRMSE ($r = 0.92–0.97$, $p < 0.01$), and strong to very strong correlation in respect of R^2 ($r = 0.79–0.93$, $p < 0.01$).

The correlation between the prevalence and AUC values were negative for all methods, whereas correlation between prevalence and both NRMSE and R^2 were

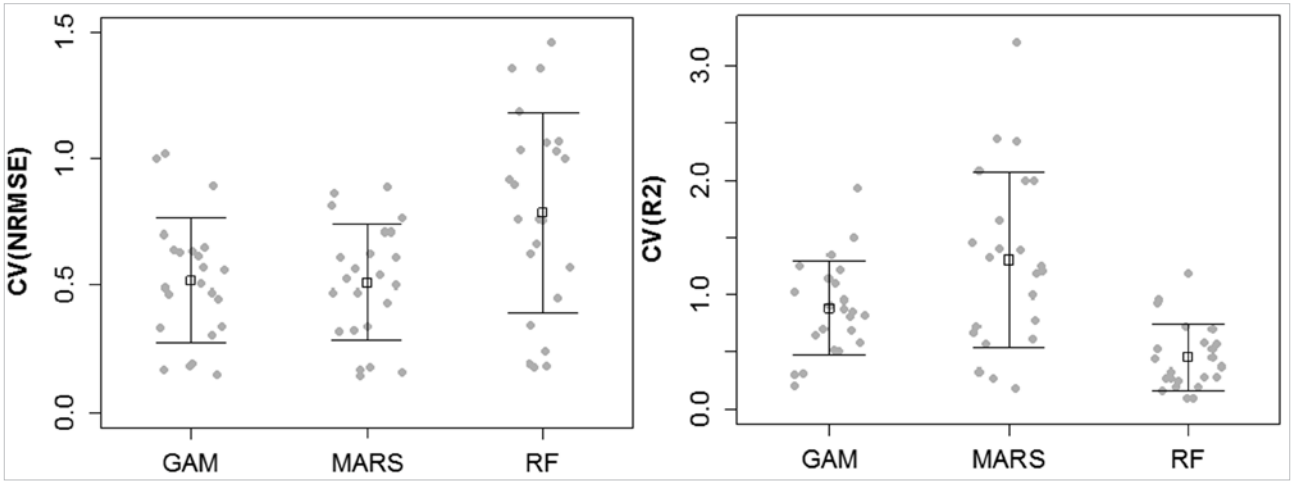

Fig. 4 Coefficients of variation of NRMSE (root mean square error normalized by range) and R^2 (coefficient of determination) of four predictive modelling methods during 100 iterations. Compiled by A. Šiaulys and M. Bučas, 2012.

Table 1 A correlation matrix of the performance of four predictive modelling methods and species prevalence. Compiled by A. Šiaulys, 2012.

	AUC				NRMSE			R^2		
	GAM	MARS	RF	MaxEnt	GAM	MARS	RF	GAM	MARS	RF
GAM	1				1			1		
MARS	0.76	1			0.97	1		0.93	1	
RF	0.90	0.76	1		0.92	0.92	1	0.81	0.79	1
MaxEnt	0.66	0.39	0.75	1						
Prevalence	-0.41	-0.13	-0.51	-0.92	0.47	0.45	0.45	0.15	0.23	0.20

AUC – area under the curve, NRMSE – root mean square error normalized by range, R^2 – coefficient of determination.

Table 2 Correlation matrix of the variation of performance of four predictive modelling methods and species prevalence. Compiled by A. Šiaulys, 2012.

	AUC				NRMSE			R^2		
	CV_{GAM}	CV_{MARS}	CV_{RF}	CV_{MaxEnt}	CV_{GAM}	CV_{MARS}	CV_{RF}	CV_{GAM}	CV_{MARS}	CV_{RF}
CV_{GAM}	1				1			1		
CV_{MARS}	0.77	1			0.95	1		0.70	1	
CV_{RF}	0.76	0.62	1		0.71	0.73	1	0.59	0.82	1
CV_{MaxEnt}	0.34	0.12	0.68	1						
Prevalence	-0.17	-0.46	-0.01	0.09	-0.41	-0.43	-0.51	-0.22	-0.48	-0.46

AUC – area under the curve, NRMSE – root mean square error normalized by range, R^2 – coefficient of determination.

always positive (Table 1). This indicates that methods tend to predict occurrence better with less occasions of species presence. While this effect was very weak for MARS ($r = -0.13$, $p > 0.05$) and moderate for GAM and RF ($r = -0.41$ and $r = -0.51$ respectively, $p < 0.05$), MaxEnt models had a very strong negative correlation with prevalence. On the contrary, performance of biomass distribution models tend to get better with increasing prevalence, however this effect was moderate in case of NRMSE and only weak-very weak in case of R^2.

Low to moderate negative correlations ($r \leq -0.51$) were determined between prevalence and coefficients of variation of models except MaxEnt (Table 2), meaning that consistency of predictions during iterations increases with decreasing occasions of species occurrence. MARS was the most sensitive in case of occurrence models, whereas biomass distribution models showed similar results for all SDMs.

Modelling results of all 23 species are given in Table 3. RF models of occurrence achieved top performance among all methods for 19 species, followed by GAM (six species), MARS and MaxEnt (one species per each). MaxEnt and MARS showed the worst performance for 14 and 10 species, respectively, whereas for GAM and RF that was never the case. According coefficient of determination RF was the best in predicting biomass distribution for all species except one, while GAM and MARS had the worst predictive performance for 14 and 12 species, respectively. Overall predictions of the occurrence were excellent ($AUC > 0.9$) for eight species: *Halicryptus spinulosus, Mytilus edulis, Saduria entomon, Fabricia sabela, Idotea balthica, Jaera albifrons,* Ostracoda, *Pontoporeia affinis* and *Theodoxus fluviatilis*, however only ostracods were predicted excellent by all four methods. The most accurate predictions for the distribution of biomass were recorded for *Balanus improvisus, Macoma balthica, M. edulis* and *S. entomon* ($R^2 > 0.5$).

DISCUSSION

All predictive modelling techniques provided useful models. In accordance with other studies (Gislason *et al.* 2006; Cutler *et al.* 2007; Collin *et al.* 2011), in our case the machine learning RF method achieved the best predictive performance on both occurrence and biomass data. However the predictive performance of models by GAM, MARS and MaxEnt was close to RF. The performance of RF models were relatively "good" ($AUC > 0.8$) for most of the species, while other modelling methods, especially MaxEnt, were "good" only for few species. The MaxEnt case can be explained by a very strong negative correlation between AUC and prevalence (Table 3), indicating that MaxEnt was relatively inaccurate for the widespread species ($AUC = 0.56–0.61$, prevalence ≥ 0.68), such as bivalve *Macoma balthica* and polychaete worms: *Marenzelleria neglecta, Hediste diversicolor* and *Pygospio elegans*. On the other hand the predictive performance of MaxEnt and other methods were relatively good for less dispersed species ($AUC = 0.90–0.93$, prevalence ≤ 0.14), such as coastal hard-bottom associated *Fabricia sabela, Idotea balthica* and *Theodoxus fluviatilis*. For the biomass models RF was superior over GAM and MARS. In most of the RF models the coefficient of determination explained up to 40% of variance more than the GAM and MARS models. This was most notable in *Mya arenaria* and *Bathyporeia pilosa* models, where the coefficients of determination of GAM and MARS were < 10%, while RF achieved > 40%.

Table 3 Validation results of four predictive modelling methods for occurrence and biomass distribution of 23 benthic species in Lithuanian economic zone. Compiled by A. Šiaulys, 2012.

Phylum, class, *order, family*	Species, taxa	GAM			MARS			RF			MaxEnt	Prev.
		AUC	NRMSE	R^2	AUC	NRMSE	R^2	AUC	NRMSE	R^2	AUC	
Priapulida	*Halicryptus spinulosus*	0.89	0.08	0.19	0.88	0.06	0.32	0.91	0.06	0.46	0.79	0.33
Polychaeta												
Nereidae	*Hediste diversicolor*	0.76	0.15	0.08	0.79	0.15	0.11	0.79	0.15	0.41	0.60	0.68
Polynoidae	*Harmothoe sarsi*	0.80	0.07	0.01	0.74	0.08	0.01	0.79	0.06	0.14	0.75	0.19
Sabellariidae	*Fabricia sabela*	0.92	0.07	0.07	0.84	0.08	0.11	0.95	0.06	0.19	0.91	0.10
Spionidae	*Marenzelleria neglecta*	0.80	0.10	0.15	0.77	0.10	0.12	0.81	0.10	0.28	0.56	0.81
	Pygospio elegans	0.83	0.08	0.01	0.79	0.08	0.01	0.80	0.08	0.14	0.61	0.69
	Streblospio shrubsolii	0.87	0.05	0.03	0.78	0.05	0.03	0.87	0.06	0.11	0.83	0.07
Oligochaeta	Oligochaeta undet.	0.77	0.04	0.05	0.77	0.05	0.01	0.82	0.04	0.08	0.65	0.59
Crustacea												
Amphipoda	*Bathyporeia pilosa*	0.82	0.05	0.06	0.79	0.05	0.02	0.86	0.04	0.46	0.79	0.22
	Corophium volutator	0.84	0.10	0.07	0.84	0.09	0.17	0.86	0.04	0.25	0.73	0.39
	Gammarus spp.	0.81	0.09	0.13	0.76	0.08	0.15	0.83	0.09	0.45	0.75	0.29
	Pontoporeia affinis	0.88	0.08	0.06	0.83	0.09	0.03	0.90	0.04	0.18	0.80	0.15
Cirripediae	*Balanus improvisus*	0.89	0.06	0.31	0.81	0.05	0.30	0.89	0.05	0.55	0.84	0.22
Isopoda	*Idotea balthica*	0.79	0.07	0.06	0.77	0.07	0.01	0.87	0.06	0.04	0.93	0.03
	Jaera albifrons	0.93	0.06	0.10	0.78	0.06	0.03	0.93	0.04	0.36	0.88	0.15
	Saduria entomon	0.92	0.14	0.32	0.92	0.15	0.35	0.94	0.15	0.71	0.83	0.31
Ostracoda	Ostracoda undet.	0.95	0.06	0.15	0.95	0.07	0.07	0.96	0.05	0.38	0.92	0.10
Gastropoda												
Hydrobiidae	*Hydrobia sp.*	0.76	0.09	0.02	0.75	0.09	0.04	0.78	0.09	0.13	0.70	0.38
Neritidae	*Theodoxus fluviatilis*	0.93	0.07	0.09	0.82	0.06	0.01	0.91	0.05	0.23	0.90	0.14
Bivalvia												
Cardiidae	*Cerastoderma lamarcki*	0.74	0.07	0.06	0.70	0.06	0.07	0.80	0.05	0.17	0.73	0.13
Myidae	*Mya arenaria*	0.82	0.07	0.02	0.80	0.07	0.06	0.89	0.05	0.43	0.69	0.48
Mytildae	*Mytilus edulis*	0.86	0.15	0.38	0.80	0.15	0.42	0.90	0.17	0.55	0.82	0.25
Tellinidae	*Macoma balthica*	0.84	0.15	0.40	0.80	0.15	0.43	0.85	0.15	0.63	0.59	0.76

AUC – area under the curve, NRMSE – root mean square error normalized by range, R^2 – coefficient of determination, Prev. – prevalence.

Our study showed that the splitting of the data into train and test datasets can play a significant role on the performance of the models. Depending on how the data was split the models performed from "poor" (AUC < 0.7) to "perfect" (AUC > 0.9). In this respect the most sensitive method was MARS for both the occurrence and biomass distribution models. For example, on the average of 100 iterations, the models of *M. balthica* achieved a "good" prediction performance, but iterations ranged from the accuracy of a coin-flip to perfect (AUC = 0.56–0.94). The most consistent method was again RF in respect of AUC and R^2.

In this study the same set of predictors was used for all the species. This was done to be more consistent when comparing different modelling techniques. However, because these species are associated to different environments, due to a limited and rigid set of parameters some of them were modelled inaccurately. For example a model performance of small isopods *I. balthica* should increase significantly if the data on the macroalgae, to which they are associated (Vetter *et al.* 1999), would be available and included. Similarly, the data on the total organic content is very important (Gogina, Zettler 2010), especially for predictions of deposit feeders such as polychaetes *H. diversicolor, M. neglecta and P. elegans* (Olenin 1997). On the other hand uneven distribution of sampling sites can result in different spatial accuracy (Šiaulys *et al.* 2012), thus in our case more dense sampling in deeper areas should provide better models for widespread and deep living species. However, both occurrence and biomass distribution models of several species were relatively good, i.e. *M. balthica*, *Mytilus edulis*, *Saduria entomon* and can provide reliable spatial maps (Fig. 5) for further ecological studies or marine spatial planning and management.

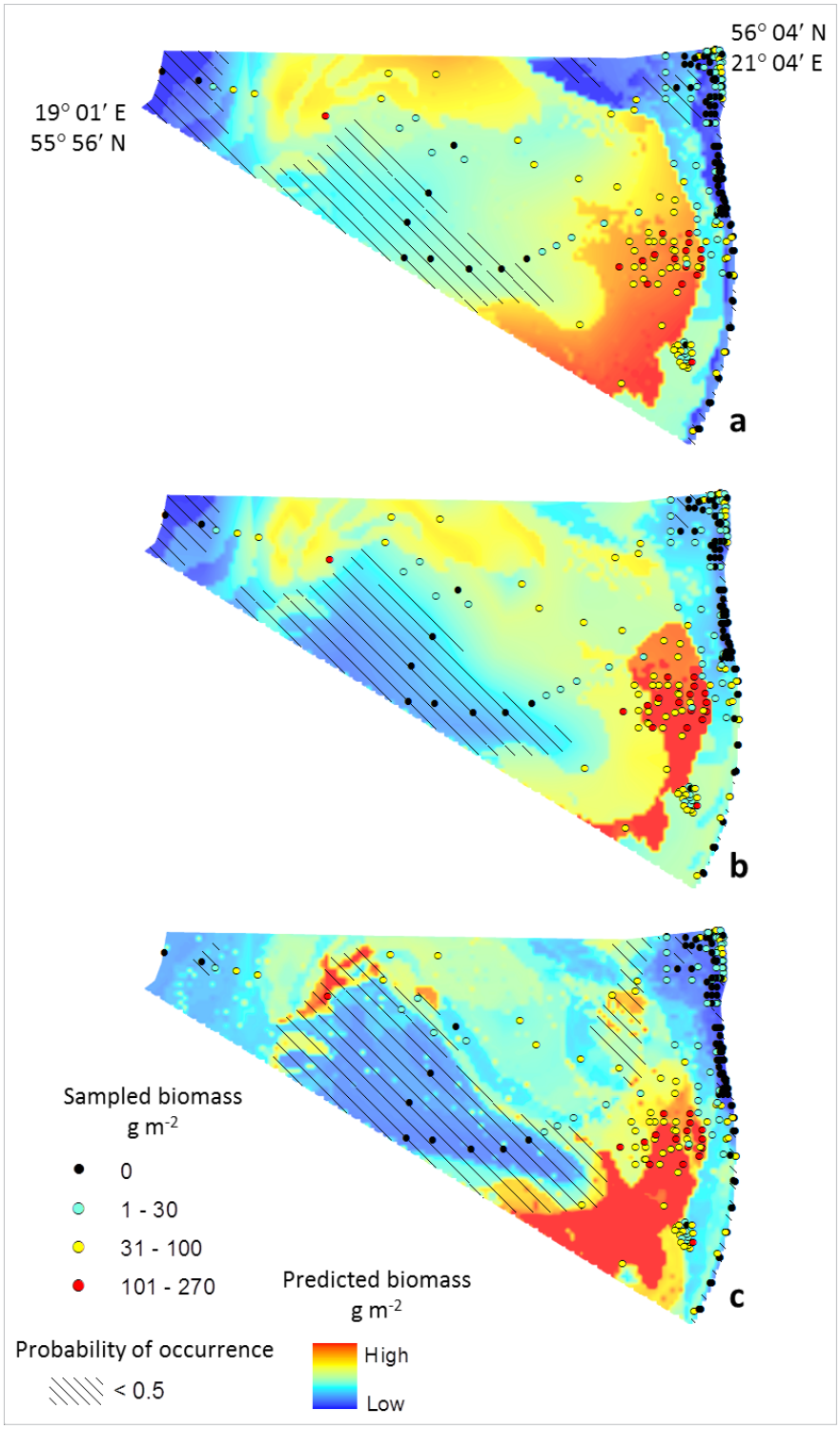

Fig. 5 Prediction graphs of *Macoma balthica* occurrence and biomass distribution modelled by different techniques: (a) generalized additive model (GAM); (b) multivariate adaptive regression splines (MARS); and (c) random forests (RF). Compiled by A. Šiaulys, 2012.

CONCLUSIONS

The RF method showed the best results in predicting the occurrence and biomass distribution of benthic invertebrates and was most consistent in relation to the data splitting into train and test datasets. Predictive performance of GAMs and MARS followed RF, whereas MaxEnt accurately predicted occurrence only for the species with relatively low distribution range.

Acknowledgements

The authors are grateful to paper reviewers Professor Jan Marcin Węsławski (Gdańsk) and Dr. Mats Lindegarth (Goeteborg) for their constructive comments, and to K. D. Fowler (Dunedin, New Zealand) for language assistance. This study was supported by BONUS PREHAB and Norwegian Financial Mechanism (Project No. LT0047).

References

Allnutt, T.F., McClanahan, T.R., Andréfouët, S., Baker, M., Lagabrielle, E., McClennen, C., Rakotomanjaka, A.J.M., Tianarisoa, T.F., Watson, R., Kremen, C., 2012. Comparison of marine spatial planning methods in Madagascar demonstrates value of alternative approaches. Plos One 7 (2). http://dx.doi.org/10.1371/journal.pone.0028969

Bendtsen, J., Söderkvist, J., Dahl, K., Hansen, J.L.S., Reker, J., 2007. Model simulations of blue corridors in the Baltic Sea. BALANCE Interim Report No. 9, 26 pp.

Bitinas, A., Aleksa, P., Damušytė, A., Gulbinskas, S., Jarmalavičius, D., Kuzavinis, M., Minkevičius, V., Pupienis, D., Trimonis, E., Šečkus, R., Žaromskis, R., Žilinskas, G., 2004. Geological atlas of Lithuanian coasts, Baltic Sea. Lietuvos geologijos tarnyba, Vilnius, 95 pp. [In Lithuanian].

Booij, N., Ris, R.C., Holthuijsen, L.H., 1999. A third-generation wave model for coastal regions. 1. Model description and validation. Journal of Geophysical Research 104 (C4), 7649–7666. http://dx.doi.org/10.1029/98JC02622

Breiman, L., 2001. Random Forests. Machine Learning 45 (5), 32 pp.

Bučas, M., Daunys, D., Olenin, S., 2009. Recent distribution and stock assessment of the red alga Furcellaria lumbricalis on an exposed Baltic Sea coast: combined use of field survey and modelling methods. Oceanologia 51 (3), 1–19. http://dx.doi.org/10.5697/oc.51-3.359

Carlström, J., Florén, K., Isaeus, M., Nikolopoulos, A., Carlén, I., Hallberg, O., Gezelius, L., Siljeholm, E., Edlund, J., Notini, S., Hammersland, J., Lindblad, C., Wiberg, P., Årnfelt, E., 2010. Modelling of Östergötland marine habitats and natural values. County Administrative Board of Östergötland, Report No. 2010: 9, 147 pp. [In Swedish].

Collin, A., Archambault, P., Long, B., 2011. Predicting species diversity of benthic communities within turbid nearshore using full-waveform bathymetric LiDAR and Machine Learners. Plos One 6 (6). http://dx.doi.org/10.1371/journal.pone.0021265

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792. http://dx.doi.org/10.1890/07-0539.1

Elith, J., Phillips, J., Hastie, T., Dudik, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17, 43–57. http://dx.doi.org/10.1111/j.1472-4642.2010.00725.x

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura,M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Sobero´n, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species‘ distributions from occurrence data. Ecography 29, 129–151. http://dx.doi.org/10.1111/j.2006.0906-7590.04596.x

Foley, M.M., Halpern, B.S., Micheli, F., Armsby, M.H., Caldwell, M.R.,Crain, C.M., Prahler, E., Rohr, N., Sivas, D., Beck, M.W., Carr, M.H., Crowder, L.B., Duffy, J.E., Hacker, S.D., McLeod, K.L., Palumbi, S.R., Peterson, C.H., Regan, H.M., Ruckelshaus, M.H., Sandifer, P.A., Steneck, R.S., 2010. Guiding ecological principles for marine spatial planning. Marine Policy 34 (5), 955–966. http://dx.doi.org/10.1016/j.marpol.2010.02.001

Friedman, J.H., 1991. Multivariate adaptive regression splines. Annals of Statistics 19, 1–67. http://dx.doi.org/10.1214/aos/1176347963

Gelumbauskaitė, L.Ž., 2009. Character of sea level changes in the subsiding South-Eastern Baltic Sea during Late Quaternary. Baltica22 (1), 23–36.

Gelumbauskaitė, L.Ž., Grigelis, A., Cato, I., Repečka, M., Kjellin, B., 1999. Bottom sediment maps of the central Baltic Sea, Scale 1:500,000. A short description, LGT Series of Marine Geological Maps No. 1, SGU Series of Geological Maps Ba No. 54, Vilnius-Uppsala.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random Forests for land cover classification. Pattern Recognition Letters 27 (4), 294–300. http://dx.doi.org/10.1016/j.patrec.2005.08.011

Gogina, M., Zettler, M. L., 2010. Diversity and distribution of benthic macrofauna in the Baltic Sea: data inventory and its use for species distribution modelling and prediction. Journal of Sea Research 64, 313–321. http://dx.doi.org/10.1016/j.seares.2010.04.005

Guerry, A.D., Ruckelshaus, M.H., Arkema, K.K., Rernhardt, J.R., Guannel, G., Kim, C., Marsk, M., Papenfus, M., Toft, J.E., Verutes, G., Wood, S.A., Beck, M., Chan, F., Chan, K.M.A., Gelfenbaum, G., Gold, B.D., Halpern, B.S., Labiosa, W.B., Lester, S.E., Levin, P.S., McField, M., Pinsky, M.L., Plummer, M., Polasky, S., Ruggiero, P., Sutherland, D.A., Tallis, H., Day, A., Spencer, J., 2012. Modeling benefits from nature: using ecosystem services to inform coastal and marine spatial planning. International Journal of Biodiversity Science, Ecosystem Services and Management 8 (1–2), 107–121.

Guisan, A., Edwards, T.C. Jr, Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157, 89–100. http://dx.doi.org/10.1016/S0304-3800(02)00204-1

Guisan, A., Zimmerman, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147–186. http://dx.doi.org/10.1016/S0304-3800(00)00354-9

Hansen, I.S., Keul, N., Sorensen, J.T., Erichsen, A., Andersen, J.H., 2007. Baltic Sea Oxygen Maps. BALANCE Interim Report no. 17, 41 pp.

Hansen, M.H., Kooperberg, C., 2002. Spline Adaptation in Extended Linear Models (with discussion). Statistical Science 17, 2–51. http://dx.doi.org/10.1214/ss/1023798997

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

Hosmer, D.W, Lemeshow, S., 2000. Applied logistic regression. Wiley Interscience, New York, 307pp. http://dx.doi.org/10.1002/0471722146

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

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2 (3), 18–22.

Li, J., Heap, A.D., 2008. A review of spatial interpolation methods for environmental scientists. Geoscience Australia, Canberra, 137 pp.

Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38, 921–931. http://dx.doi.org/10.1046/j.1365-2664.2001.00647.x

Milborrow, S., 2012. Multivariate adaptive regression spline models. Package “Earth”, 43 pp.

Phillips, S.J., Dudik, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175. http://dx.doi.org/10.1111/j.0906-7590.2008.5203.x

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190, 231–259. http://dx.doi.org/10.1016/j.ecolmodel.2005.03.026

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. http://dx.doi.org/10.1007/s10021-005-0054-1

Olenin, S., 1997. Benthic zonation of the Eastern Gotland Basin. Netherlands Journal of Aquatic Ecology, 30 (4), 265–282. http://dx.doi.org/10.1007/BF02085871

Quantum GIS Development Team, 2010. Quantum GIS Geographic Information System. Open Source Geospatial Foundation Project; http://qgis.osgeo.org.

Quinn, G.P., Keough, M. J., 2002. Experimental design and data analysis for biologists. Cambridge University Press, Cambridge, 520 pp. http://dx.doi.org/10.1017/CBO9780511806384

Reiss, H., Cunze, S., König, K., Neumann, H., Kröncke, I., 2011. Species distribution modelling of marine benthos: a North Sea case study. Marine Ecology Progress Series 442, 71–86. http://dx.doi.org/10.3354/meps09391

Robinson, L., Elith, J., Hobday, A., Pearson, R.G., Kendall, B.E., Possingham, H.P., Richardson, A.J., 2011. Pushing the limits in marine species distribution modeling: lessons from the land present challenges and opportunities. Global Ecology and Biogeography 20, 789–802. http://dx.doi.org/10.1111/j.1466-8238.2010.00636.x

Šiaulys, A., Daunys, D., Bučas, M., Bacevičius, E., 2012. Mapping an ecosystem service: a quantitative approach to derive fish feeding grounds. Oceanologia 54 (3), 491–505. http://dx.doi.org/10.5697/oc.54-3.491

Vetter, R.A.H., Franke, H.D., Buchholz, F., 1999. Habitat-related differences in the responses to oxygen deficiencies in Idotea baltica and Idotea emarginata (Isopoda, Crustacea). Journal of Experimental Marine Biology and Ecology 239, 259–272. http://dx.doi.org/10.1016/S0022-0981(99)00049-0

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, 1471–1478. http://dx.doi.org/10.1016/j.ecolmodel.2011.02.007

Wei, C. L., Rowe, G.T., Escobar-Briones, E., Boetius, A., Soltwedel, T., Caley, M.J., Soliman, Y., Huettmann, F., Qu, F., Yu, Z., Pitcher, C.R., Haedrich, R.L., Wicksten, M.K., Rex, M.A., Baguley, J.G., Sharma, J., Danovaro, R., MacDonald, I.R., Nunnally, C.C., Deming, J.W., Montagna, P., Lévesque, M., Weslawski, J.M., Wlodarska-Kowalczuk, M., Ingole, B.S., Bett, B.J., Billett, D.S.M., Yool, A., Bluhm, B.A., Iken, K., Narayanaswamy, B.E., 2010. Global patterns and predictions of seafloor biomass using random forests, Plos One 5 (12). http://dx.doi.org/10.1371/journal.pone.0015323

Wentworth, C. K., 1922. A scale grade and class terms for clastic sediments, Journal of Geology 30, 377–392. http://dx.doi.org/10.1086/622910

Wood, S.N., 2006. Generalized additive models: An introduction with R. Chapman and Hall/CRC, 410 pp.

Wood, S.N., Augustin, N.H., 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecological Modelling 157, 157–177. http://dx.doi.org/10.1016/S0304-3800(02)00193-X