

Mapping Reef forming North Sea Species



Mapping Reef forming North Sea Species

Author(s)

P.M.J. Herman PhD

F.F. van Rees MSc

Partners

Ministerie van Infrastructuur en Waterstaat Inspectie Leefomgeving en Transport, 'S-GRAVENHAGE

G.C.S. Hommel

Mapping Reef forming North Sea Species




Client	Ministerie van Infrastructuur en Waterstaat
Contact	G.C.S. Hommel
Reference	
Keywords	MONS, Habitat suitability, <i>Sabellaria spinulosa</i> , <i>Modiolus modiolus</i> , <i>Lanice conchilega</i> , <i>Ostrea edulis</i>

Document control

Version	1.0
Date	25-02-2022
Project nr.	11207716-000
Document ID	11207716-000-ZKS-0002
Pages	40
Classification	
Status	final

Author(s)

	P.M.J. Herman	
	F.F. van Rees	

Doc. version	Author	Reviewer	Approver	Publish
1.1	P.M.J. Herman 	L.A. van Duren 	P. Visman Paul Saager 	
	F.F. van Rees			

Summary

The aim of this project was to predict as accurately as possible where four reef-building species in the North Sea: *Sabellaria spinulosa*, *Modiolus modiolus*, *Lanice conchilega* and *Ostrea edulis* can develop stable populations, given the environmental gradients and the gradients in human use of the North Sea. This report documents the compilation of a habitat suitability map.

For the three species that currently have stable populations in the North Sea (*Sabellaria*, *Modiolus* and *Lanice*), spatial distribution data were derived from EMODnet Biology. In addition, for *Sabellaria* and *Modiolus* data of macrobenthic bycatch in fishing trawls of Wageningen Marine Research were used. *Lanice* was too sparsely distributed in these data to be considered reliable. For *Ostrea edulis*, historic data on the distribution in the nineteenth and early twentieth century in Dutch, German and Belgian waters was used from literature sources. Environmental data were derived from a recent compilation in the literature, and in addition from Deltares physical modelling.

The presence/absence of the species was regressed on the environmental data using two regression models: logistic regression and random forest regression. The results of both models were in close agreement for all species. The random forest models gave the finest-grained predictions and are taken as the final product of the project.

In this document, the R code used for collecting all data, preparing GIS files of the data and performing the regression analyses is discussed. The present report is accompanied by a QGIS project and all its underlying files, as well as by a geo-pdf file that contains all the environmental layers, species occurrence data and regression predictions

Contents

	Summary	4
1	Introduction	6
2	Documentation of data procedures	8
2.1	Structure of the project	8
2.2	Preliminary settings	8
2.3	Prepare all environmental information	9
2.4	Extracting species presence/absence information from the available data sets	10
2.5	Retrieving data from Wageningen Marine Research fisheries database	10
2.6	Retrieving historical data on oyster distribution	10
2.7	Collecting all species information and linking to environment	11
2.8	Visualizing species environment relations	11
2.9	Regression analysis	11
3	Results and Discussion	13
3.1	Sabellaria spinulosa	14
3.2	Modiolus modiolus	16
3.3	Lanice conchilega	17
3.4	Ostrea edulis	18
4	Acknowledgements	20
5	References	21
A	Appendices	22
A.1	R code for the analysis	22
A.1.1	Code chunk#1	22
A.1.2	Code chunk #2	22
A.1.3	Code chunk #3	25
A.1.4	Code chunk #4	25
A.1.5	Code chunk #5	26
A.1.6	Code chunk #6	28
A.1.7	Code chunk #7	28
A.1.8	Code chunk #7	29
A.1.9	Code chunk #9	30
A.2	Exploratory species-environment plots	31

1 Introduction

The purpose of the analysis described in this document is to collect and interpret data on the spatial distribution of reef-forming species that can potentially occur in Dutch North Sea waters. This is part of the MONS research programme. It is executed by Deltares on demand of Rijkswaterstaat.

The species of interest for this analysis are *Sabellaria spinulosa*, *Modiolus modiolus*, *Lanice conchilega*, *Ostrea edulis*. All four of these species can form biogenic reefs, which in turn can form hotspots for biodiversity as they provide attachment or hiding opportunities for other species. Biogenic reefs are recognized as habitats worthy of special protection in OSPAR and EU regulations. That applies in particular to *Sabellaria spinulosa* and *Modiolus modiolus*, where trawling is considered to endanger the special habitats created by the reefs. *Lanice conchilega* is a very common species that does not appear to be particularly threatened, whereas *Ostrea edulis* is practically extinct in the North Sea, but is currently the subject of intense restoration efforts.

The aim of the current analysis is to predict as accurately as possible where the species of interest can develop stable populations, given the environmental gradients and the gradients in human use of the North Sea. For the three species that are currently occurring in the North Sea, the habitat preference is deduced from their current occurrence patterns and the spatial distribution of relevant environmental characteristics. Fisheries intensity was used as a co-factor in these analyses, in an attempt to delineate how fisheries pressure by different gear affects the current occurrence patterns. For *Ostrea edulis*, this approach was not possible as the species is currently not recorded in the North Sea, with the exception of some recent population developments at the very margin of the North Sea in Voordelta, Rotterdam harbour and Wadden Sea, and occurrences on buoys and other artificial structures. The spatial coordinates of the latter sparse observations cannot be used as an indicator of habitat suitability for stable oyster populations, as they are dependent on artificial substrates. Physiologically, flat oysters can probably survive in most of the North Sea provided a larval source and hard substrate are available. This is, however, not indicative of the range of habitats where the species might reestablish a stable, self-sustaining population. In order to deduce the spatial delineation of potentially suitable areas for restoration of flat oysters, data on historical abundance in Dutch and Belgian waters have been used, based on Bennema et al. (2020) and Houziaux et al. (2008). It is possible that some historic occurrences along the British coasts have been missed by this selection, but for the Dutch EEZ and its immediate surroundings, this selection may suffice.

Available occurrence data for the three other species have recently been compiled in the framework of EMODnet Biology (Herman et al., 2020). By carefully selecting the data sets that have, in principle, looked for the entire macrobenthic community (or for a well-defined part thereof, e.g. all shellfish), the presence-only database has been transformed into a presence/absence dataset. It has been assumed that wherever a sample targeting the entire macrobenthic community has taken, all macrobenthic species not recorded in the sample were actually absent in the sample. Therefore, all these species have been attributed an 'absence' record in all community samples where they have not been found. In total, more than 60 data sets covering almost 100,000 samples have been collected in the Greater North Sea, which also includes the Irish sea and part of the N.E. Atlantic. the number of samples in the North Sea proper is around 20,000. In addition to this dataset, WMR has made available the data from all the fish surveys they have performed in the North Sea. In the fish surveys, bycatch of benthic animals is recorded. Especially for *Sabellaria spinulosa* and *Modiolus modiolus* this results in regular reporting of presence of the species. The number of positive recordings for *Lanice conchilega* was so low that we estimated the catchability of this species by the fishing gear is not sufficient to use the data base for their distribution. *Ostrea edulis* was not reported in this data set. For the two species of interest, all samples where the species was not recorded, was noted as 'absence' of the species. These data were then added to the EMODnet data base.

Here we make use of the presence/absence data in the North Sea proper, thus neglecting data in the Channel and Irish Sea. The reason for the selection is that for the North Sea proper we can make use of environmental data collected and made available by van der Reijden et al. (2018). By regressing the presence/absence data on the environmental data set, we can gain some insight in the environmental parameters steering the spatial distribution of the species, but we can also refine and improve the spatial interpolation between observations. In this analysis, we applied two regression techniques: logistic regression and random forest regression. Results of both approaches are given. These results compare favourably, thus providing credibility to the estimated patterns of occurrence.

The present document is set up as an R Markdown document. The Latex code of the document is generated while the enclosed R code is executed, thus guaranteeing simultaneous and consistent execution of the code documented in the text. In the document we show the scripts for the different steps in the analysis.

2 Documentation of data procedures

2.1 Structure of the project

The R script in this document analyses the underlying data sets and produces a number of ESRI shapefiles (for vector data) and geotiff raster files (for raster data) out of these data sets. All these GIS files have been compiled into a QGIS project that can be found back in the underlying directory structure of the project. In addition, all layers have been exported from the GIS project into a geo-pdf file, which allows to show all layers independently or in combinations within Acrobat Reader, and which should also allow to retrieve all layers in a GIS, without recourse to the original shape and raster files.

The directory structure of the project is shown in Table 2.1.

Table 2.1 Directory structure of the project

Directory	Content
project root directory	
-base_data	data to be read in by the scripts and used in the products
.. -EMODnet_data	data from the EMODnet data base. These include the Dutch MWTL data
.. -data_WMR	data provided by WMR. Only data from the fish surveys (Datras) have been used
.. -Ostrea	data from Bennema et al. (2020) and Houziaux et al. (2008)
-Environment	data on environmental factors
.. -DCSM-FM	output of the Deltares North Sea model, used for bottom shear stress
.. -Environmental_factors	raster files provided by van der Reijden et al.(2018)
.. -Fisheries_data	data on fisheries intensity by van der Reijden et al.(2018)
.. -rasters	resampled rasters of environmental factors, used in analysis
-Europe_coastline_shapefile	downloaded from EEA, used to blank the rasters over land
-output	rasters with the predicted values for all species from the two models
-QGIS	QGIS project file
-Shapefiles_species	Shapefiles with the presence/absence data per species, except for the DATRAS data
.. -WMR	Shapefiles with the DATRAS data for the species

2.2 Preliminary settings

Code chunk #1 in the appendix specifies the R packages loaded to execute the analyses. It further sets some constants such as directory names and projection strings. All analysis of spatial data will use UTM zone31 coordinates.

2.3 Prepare all environmental information

Environmental information is needed as a basis for species distribution models. For this project, we rely heavily on a recent compilation of North Sea wide environmental information by van der Reijden et al. (2018). These authors have compiled their datasets on bathymetry, grain size distribution, temperature and salinity from diverse literature sources. They have made their data available in the form of geo-tiff files, that we have downloaded for use in the present project. In the files, there is also information on bottom shear stress, but this is based on a rather coarse model. We have replaced it with results of the Deltares DCSM-FM model for the greater North Sea. The datasets used are listed in Table II. Sources of the data are van der Reijden et al. (2018) for fisheries and calculations of 'Bathymetric Position Index' values based on bathymetry, Stephens (2015) for grain size data, Copernicus marine services ([www.marine.copernicus.eu](http://portal.emodnet-bathymetry.eu/)) for salinity and temperature, EMODnet bathymetry (<http://portal.emodnet-bathymetry.eu/>) for basic bathymetry, Deltares for bottom shear stress calculated with DCSM-FM.

The 'BPI' (Bathymetric position index) calculates for each point, the difference of the depth of the point with the average depth of the surrounding area, where the surrounding area is a circle with a fixed radius. BPI5 uses 5 km as a radius for the surroundings, and similar for the other BPI variables. van der Reijden et al. (2018) also define a weighted average BPI, but we did not use that in our analysis.

Temperature difference is a measure for the change in temperature between 2008 and 2013. This is not distributed homogeneously over the North Sea. Atlantic water has warmed very little, whereas the North Sea has been warming considerably over the past decades. Consequently, the largest temperature differences are seen in the eastern and north-eastern parts of the North Sea.

No temporal (e.g. seasonal) variance of salinity and temperature has been used in the present study. It is known that variation of these variables is often very important in estuarine conditions. However, in the North Sea the ranges are much more limited. It is unlikely that any of these parameters would fall outside of the tolerance of the species, with the probable exception of temperature for the boreal species *Modiolus modiolus*. However, also mean temperature appeared to be a very useful variable in predicting the range of this species, and obviously there is a tight correlation between mean temperature and yearly temperature range in the North Sea.

Table 2.2. Environmental data and their source

Env.Variable	Explanation	Source
Depth	Depth at 178 m resolution	EMODnet
BPI5	Bathymetric Position Index 5 km	vdReijden2018
BPI10	Bathymetric Position Index 10 km	vdReijden2018
BPI75	Bathymetric Position Index 75 km	vdReijden2018
Bott.shr.stress	Bottom shear stress from currents	DCSM-FM
Salinity	Mean Salinity	Copernicus
Temperature	Mean Temperature	Copernicus
Temp.diff	Temperature Difference over the year	Copernicus
Gravel	Fraction gravel in sediment	Stephens2015
Mud	Fraction Mud in sediment	Stephens2015
Sand	Fraction Sand in sediment	Stephens2015
Beam_plaice	Intensity beam trawling for plaice	vdReijden2018
Beam_sole	Intensity beam trawling for sole	vdReijden2018
Otter_mix	Intensity otter trawling for mixed species	vdReijden2018

The code in code chunk#2 is used, first to extract and reconfigure the model results on bottom shear stress from currents, subsequently to read all environmental factors, and to

resample them to the same resolution for all variables. These intermediate rasters (gridded data) are stored for later use, as it is a quite time-consuming process.

The environmental rasters are used for two purposes. First, for every sample the environmental information can be read from the rasters. This will complete the data frame with sample information. This step is performed later in the code, as the data frame with species occurrence has first to be prepared. Secondly, the rasters are used as a basis for the predictions based on the regression models. A data frame 'newdats' is made, that contains for every point on the raster, the coordinates of the point and the values of all environmental variables in the point. Using these data and the regression model, a model prediction can be made for every point on the raster. These predictions are then again assembled in a raster and presented in GIS.

2.4 Extracting species presence/absence information from the available data sets

The EMODnet Biology product on presence/absence of species in samples in the Greater North Sea is delivered as a binary R file. Alternatively, it is also available as a .csv file, but this takes longer to read in.

In code chunk#3 a function is defined that retrieves the data for a particular species and writes the results as a shape file for use in GIS. Species are identified using their AphiaID, which is their unique identification in WoRMS, the World Register of Marine Species (<https://marinespecies.org>)

2.5 Retrieving data from Wageningen Marine Research fisheries database

Wageningen Marine Research has made available all data in their 'Frisbee' database on the concerned species (called 'DATRAS' data). The data set is composed of all hauls with a diversity of instruments, including beam trawls, otter trawls, plankton nets and others. The species concerned were never retrieved from some of these instruments, probably because some instruments (e.g. plankton nets) are not able to catch them. In order to avoid excess zeroes, suggesting absence of the species whereas presence could not have been established, we restricted the database to those instruments that had at least once caught one of the concerned species. These are beam trawls, otter trawls and an instrument called 'GOV'. Closer examination showed that *Lanice conchilega*, one of the most frequently found species of macrobenthos in the North Sea, was only found 12 times in total in this database. We concluded that inclusion of the database for this species would lead to too many false zeroes, and restricted use of the database to *Modiolus* and *Sabellaria* only. The oyster was not reported from this database. However, in the retrieval code illustrated here, all four species are looked after in the DATRAS data base and illustrating shapefiles for all four are produced.

Code chunk#4 was used to extract the data from the DATRAS database.

2.6 Retrieving historical data on oyster distribution

Historical data on the distribution of oysters in the North Sea, and more particularly in the Dutch waters, during the nineteenth century were derived from Bennema et al. (2020), and courteously made available to us by Floris Bennema. These authors discuss two different sources of data in their paper. One source are historical expeditions in the North Sea, the data of which have been digitized. We received these data in two files: one file describing finds by the Huxley_Wodan expeditions, and one by the Poseidon expeditions. These data have been read in and converted to spatial files. The other source were old maps, that have been critically evaluated by the authors and compiled into an overall map indicating the area of high oyster occurrence in the region around the Oyster Grounds. We digitized this map into a polygon using QGIS and used it as a basis to generate pseudo-absences and pseudo-presences. Random points were generated in the North Sea, and points within the map polygon were attributed a probability of 0.7 to contain oysters, whereas points outside of the

polygon had absence. In order to complement this data base with information on the Flemish Banks, that could also be of importance to the Dutch waters off Zeeland, we used the report by Houziaux et al. (2008) on the findings of the extensive set of dredge surveys by Gilson in the beginning of the twentieth century. We digitized all sample points of Gilson from the figures in the report of Houziaux, indicating presence of oysters where this was recorded. The points were saved as a shapefile and read in to extract the points with absence and presence information. The data provided by Wageningen Marine Research also contain findings of flat oysters in the Voordelta, the Rotterdam harbour and the Wadden Sea. All three of these populations fall outside of the environmental rasters available in the present project. Two of them seem to depend on artificial hard substrate, although it remains to be seen if that is only a transition phase or not. It is also likely that in these estuarine or near-estuarine conditions, other environmental factors (e.g. salinity) will have an influence on habitat suitability than in the open North Sea. For these reasons, information from these populations was not used in the present analysis, which was restricted to historical data of oyster occurrence on natural substrates.

All data manipulations regarding oysters are documented in code chunk#5.

2.7 Collecting all species information and linking to environment

Having prepared the distribution data for the four species, and all environmental information, the next step (code chunk #6) brings all of this information together. Per 'observation event' (usually a sample) the presence/absence of the four species is recorded, and the value of each of the environmental variables for the coordinates of the sample is extracted from the rasters. This file is stored and will be used in the regression analyses.

2.8 Visualizing species environment relations

As a preliminary analysis, plots are produced showing the raw data of species occurrence versus the environmental factors in the database. Observations are split in twelve groups of increasing value of the environmental variable. Each of the groups has an equal number of observations. Per group, the mean occurrence of the species in the group is plotted versus the mean value of the environmental variable in the group. Ranges of the environmental variable are also indicated. These plots are purely exploratory, in order to obtain a first visual impression of the degree of correlation between the species and the environmental factors. The code is given in code chunk#7.

Appendix A.2 gives all plots for the four species.

2.9 Regression analysis

Species distribution models have been prepared with two different regression techniques: logit regression and random forest regression. For the logit regression, the environmental variables and their squared values have both been entered into the regression equation, allowing for Gaussian-type response curves. In general, the predictions of the logit regression are smoother in space than the random forest regressions, probably because the responses on the environment are necessarily smooth in these parametric functions. Random forests, on the other hand, are based on a classification approach and can use very sharp boundaries in the environmental variables to have a strongly different effect on the modeled variable. However, apart from these relatively subtle differences, both methods give very similar predicted spatial patterns for the species. The fisheries intensity was not relevant as a predictor for the oyster, as the oyster data are historical nineteenth-century reconstructions. It turned out that for the other three species, the predictive power of the three fisheries intensities was very low. The variables have been removed from the analysis. Furthermore, sand fraction has also been removed from the analysis, because it is fully collinear with mud and gravel fractions: the three together always sum to 1. From the BPI variables, we only retained BPI at 5, 10 and 75 km, as the other classes (30 and 50 km) were usually redundant with these three. The remaining variables all had at least some importance in almost all regression models. If a single factor was occasionally not statistically relevant, it was still

maintained in the analysis in order to keep consistency between the different species. Significance of each factor can only be established for the logit regression, but even here the significance may be biased due to spatial autocorrelation. We did not attach too much importance to the calculated significance. In the random forest model, there were clear signs of overfitting in the *Ostrea* model, when only the expedition data were used. Overfitting was manifested because the prediction model only predicted occurrence in a very narrow band around the positive observations, not in between them. This defect was much less apparent after we added the pseudo-data based on the historical maps. For the other random forest models, no clear signs of overfitting were apparent, although it might sometimes be the case in the *Modiolus* map.

Code for the regression models is given in code chunks#8 and #9.

3 Results and Discussion

We used the logistic regression results mainly as a control of the random forest models. Random forest regression models are more versatile and better able to catch non-linearities in the responses. However, they are vulnerable to overfitting and may provide spurious results in data-poor areas. For this reason, we also present both models in the present report. However, we consider the random forest predictions, which are very well endorsed by the logistic regression in all four cases, as the main results of this project.

Table 3.1 summarizes the coefficients of all terms in the logistic regressions and indicates their (approximate) significance. It can be seen that for most models, the majority of the environmental variables contributed to the regression model. We did not prune the models any further, as they are primarily meant to interpolate the available data, using the environment as additional information to improve the interpolation.

Table 3.1. Summary of logistic regression models. Per species the coefficients for the different terms in the regression model are given ("Coef"), as well as their probabilities ("Pr"), coded as: *** <0.001; ** <0.01; * <0.05; . <0.1

terms	Coef Sabel	Pr Sabel	Coef Modiol	Pr Modiol	Coef Lanic	Pr Lanic	Coef Ostrea	Pr Ostrea
(Intercept)	-128.36	.	-42.23		71.10	***	-360.15	**
mean_stress	6.14	***	3.16	***	-1.23	***	-3.49	*
l(mean_stress^2)	-1.79	***	-1.21	***	0.29	***	1.33	
depth	0.11	***	0.0099		0.064	***	-0.31	***
l(depth^2)	0.0003	**	0.0001		0.0006	***	-0.0008	.
bpi5	-0.13	***	-0.13	***	0.027	*	0.18	
l(bpi5^2)	-0.0041	***	-0.0025		0.0012		0.042	**
bpi10	0.20	***	0.20	***	0.10	***	-0.082	
l(bpi10^2)	0.0051	***	0.0046	**	-0.0014	*	-0.056	**
bpi75	0.024	**	-0.068	***	-0.034	***	-0.15	***
l(bpi75^2)	-0.0027	***	-0.0023	***	-0.0010	***	-0.0048	**
meantemp	5.89	***	-0.99		-2.72	**	-0.52	
l(meantemp^2)	-0.30	***	-0.021		0.11	**	0.10	
difftemp	-0.44	.	-0.70	*	1.13	***	4.88	***
l(difftemp^2)	0.0013		0.041	**	-0.042	***	-0.19	***
salinity	5.28		1.89		-3.96	***	19.64	*
l(salinity^2)	-0.067		-0.0098		0.063	***	-0.31	**
gravel	6.68	***	4.26	**	7.76	***	12.04	
l(gravel^2)	-11.69	***	-0.60		-10.42	***	-101.24	
mud	21.88	***	-17.65	***	4.17	***	18.10	***
l(mud^2)	-70.54	***	23.71	*	-21.67	***	-42.47	***

In random forest regression, no similar quantities to 'significance' are calculated. However, there are measures of the importance of the independent variables for the model predictions. The importance is determined by comparing the full model with a submodel in which the values of one of the variables have been scrambled at random and evaluating the difference in fit of both models. That can be done on the basis of the mean square error (difference between model prediction and observation), but also with a compound goodness-of-fit variable called node impurity. The order of variables in both importance rankings was not always the same, suggesting that in most of our cases the different variables contributed rather equally to the result, without a dominant pattern emerging. This image was different when the fisheries intensities were still part of the models. In all four cases, they stood out as

extremely unimportant in the final model. For this reason, they were excluded from the analysis. We conclude that we have insufficient information (i.e. insufficient areas that are either fished or unfished but similar with respect to other environmental characteristics) to derive meaningful estimations of the effect of fisheries on the three species with contemporary data. For *Ostrea*, this analysis was excluded anyway, as the data predate the estimated fisheries effort by a century.

Table 3.2. Summary of variable importance in random forest regression models. IM: increase in MSE; INP: increase in Node Purity

terms	IM Sabel	INP Sabel	IM Modiol	INP Modiol	IM Lanice	Inp Lanice	IM Ostrea	INP Ostrea
mean_stress	93	369	39	35	142	320	28	20
depth	64	194	41	32	102	318	33	22
bpi5	97	168	36	28	130	289	19	12
bpi10	92	169	37	28	112	272	26	13
bpi75	82	181	36	30	105	284	35	15
meantemp	63	256	55	46	107	313	26	21
difftemp	63	228	50	43	123	319	27	17
salinity	76	270	46	43	131	295	33	21
gravel	96	517	48	35	156	331	53	51
mud	109	216	58	34	131	280	34	29

3.1 Sabellaria spinulosa

Two data sources have been used for this species: EMODnet data from grabs and box cores, and DATRAS fish trawl data. Both data sources have a general correspondence in the spatial pattern of occurrences, although many more positives were found in the EMODnet data set than in the trawl data. The main reason for this is that the main area of occurrence of *Sabellaria spinulosa* along the English east coast, was not heavily sampled with the fish trawls. It is probably unsuitable area for fisheries, as it is characterized by a gravel-containing and stony bottom type. The preferences of the species are clearly guided by high bottom shear stress and high gravel content of the sediment. Also, some bathymetric characteristics contribute to the pattern of expected occurrence. In the regression models, no discernible influence of fisheries intensity could be found. However, it should be stressed that no substantial unfished areas (in suitable fisheries areas) are represented in the data set. As fishermen respond strongly to environmental gradients themselves (see van der Reijden et al. 2018 for a discussion of this response), it remains difficult to establish the influence of fisheries on the distribution of animals.

The two regression models correspond fairly well in their predicted distribution pattern (Figure 3.1). The only exception is a predicted occurrence toward the Skagerrak, based on the random forest. There are, however, no observations to corroborate this prediction. Further, as is the case in all species, the random forest prediction shows sharper spatial gradients than the logistic regression. Some independent information on the occurrence of *Sabellaria spinulosa* on offshore structures was provided by WMR. These structures generally fall outside of the distribution area of *Sabellaria* as found in this study. The offshore structures may offer essential modifications of the habitat: they provide both hard substrate for attachment of the worms, and local scouring leading to high bottom shear stress and availability of mobile sand. The species is known to heavily depend on bottom load of mobile sand, as this is the resource it uses to construct the reef structures. As *Sabellaria* is relatively widespread as solitary, non-reef building individuals, it can be expected that offshore structures will harbour reef-building populations of the species, even outside of its natural region of occurrence. From the data on natural occurrence, it can be concluded that the Dutch EEZ in the North Sea is situated at or beyond the limit of the normal range of

occurrence of the species. One population near the Brown Bank is the only recorded natural occurrence of the species in Dutch waters (van der Reijden et al. 2019). The finding is relatively recent and may not reflect a sustained presence of the species, although this will have to be checked in future. Note that this finding was included in our database, but the regression model is also sensitive to the many zero observations in the surroundings, which yields a relatively low predicted chance of occurrence at the Brown Bank. In any case, although *Sabellaria* occurrence in Dutch waters is very interesting because it is at the very margin of its natural distribution, the Dutch EEZ does not seem to be the core area for protection of the species. The core of its distribution is located in highly energetic waters closer to the English coast, as well as areas in Scotland, the Channel and the Irish Sea. *Sabellaria spinulosa* is not a rare species in the Greater North Sea, occupying around position 50 (of over 4000) in the list of most frequently found macrobenthos species in the EMODnet database. Its occurrence, however, is strongly clustered.

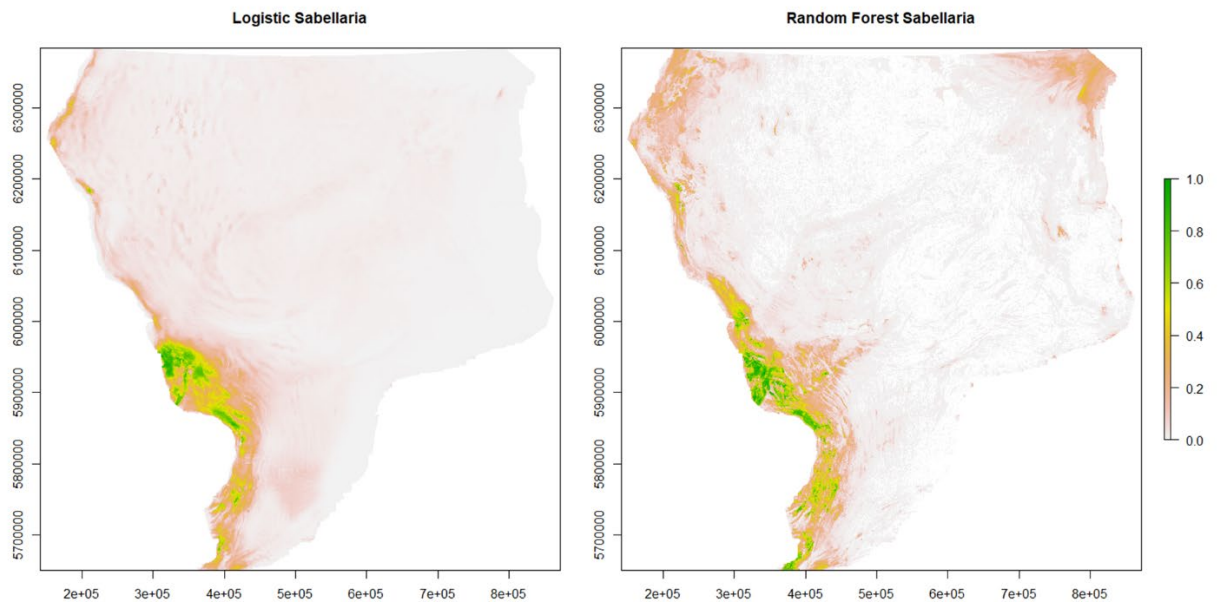


Figure 3.1. Results of the logistic (left) and random forest (right) regression models for *Sabellaria spinulosa*. The maps present the predicted probability of occurrence of the species over the entire North Sea.

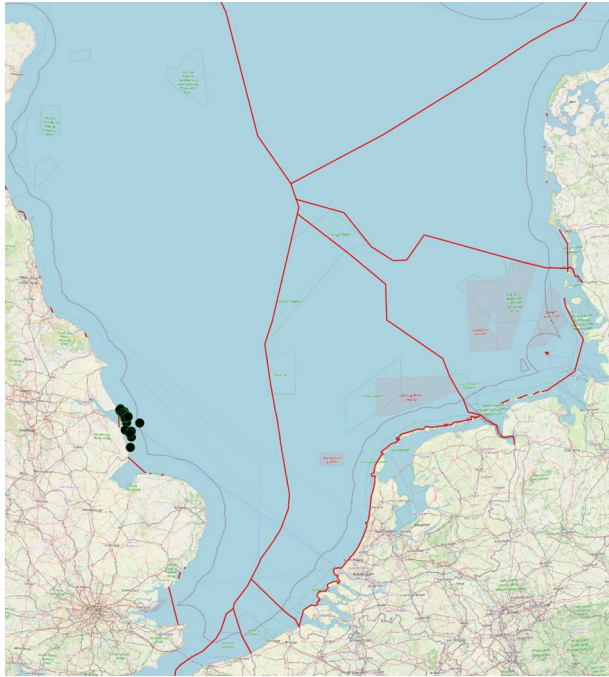


Figure 3.2 Distribution of *Sabellaria alveolata* in the EMODnet database.

3.2 *Modiolus modiolus*

The general pattern of occurrence of *Modiolus modiolus* in both regression analyses is similar: the species is restricted to the northern part of the North Sea, north of the Doggerbank (Figure 3.3). Extensive beds of *Modiolus* are known from Scottish coastal systems, on gravel, muddy gravel and bedrock substrates. The species is known to be boreal and restricted to cold waters. In the North Sea, its most southerly extensive occurrence is in gravel-rich sediments along the English coasts. The few findings along the Belgian coast, on gravely sand in between sand banks, are not reflected in the regression models. It is possible that the resolution of the environmental rasters is insufficient to reveal the sharp gradients in sediment composition or bottom shear stress that are associated with sand banks.

The species has (almost) not been recorded from the Dutch EEZ in the North Sea. A few occasional finds suggest that it may be present in low numbers, presumably attached to small stones or other hard substrate. There have been suggestions, but no hard data, that the species may have occurred on the Cleaver Bank. From its environmental preferences, this does not seem unlikely. The species is described as 'near threatened' in the EU. It has been described as 'under threat and/or declining' by OSPAR, and as 'vulnerable' by HELCOM. Although many classifications suggest a decline in its distribution, this is poorly documented and cannot be demonstrated using historical data. It is assumed that trawling is a major threat to the species. However, whether this also applies to the sandy areas where most of the Dutch trawling for sole and plaice take place, is not sure. In our data set, regression on fisheries intensities was not successful.

The two regression models coincide in their general predictions of the distribution of the species. The random forest model shows some signs of overfitting, where small patches of high occurrence probability are located tightly around the positive data points. Some caution is therefore required with the fine-grained aspects of the prediction.

The likelihood of restoring the species in the Dutch EEZ does not seem large. In most of the Dutch EEZ, the species lacks the gravel or other hard substrate that it needs for attachment. The recent rise in temperature of the water in the eastern half of the North Sea, up to almost one degree over the past 30 years, is also a very unfavourable feature for this species. It is not unlikely, however, that it may show up occasionally where artificial hard substrate is offered. The random forest model also suggests that there may be small patches, e.g. in

between sand banks, where the species might thrive. Some of these landscape features are too small-scale for our environmental rasters. Their actual number could therefore be somewhat larger than suggested by the prediction map.

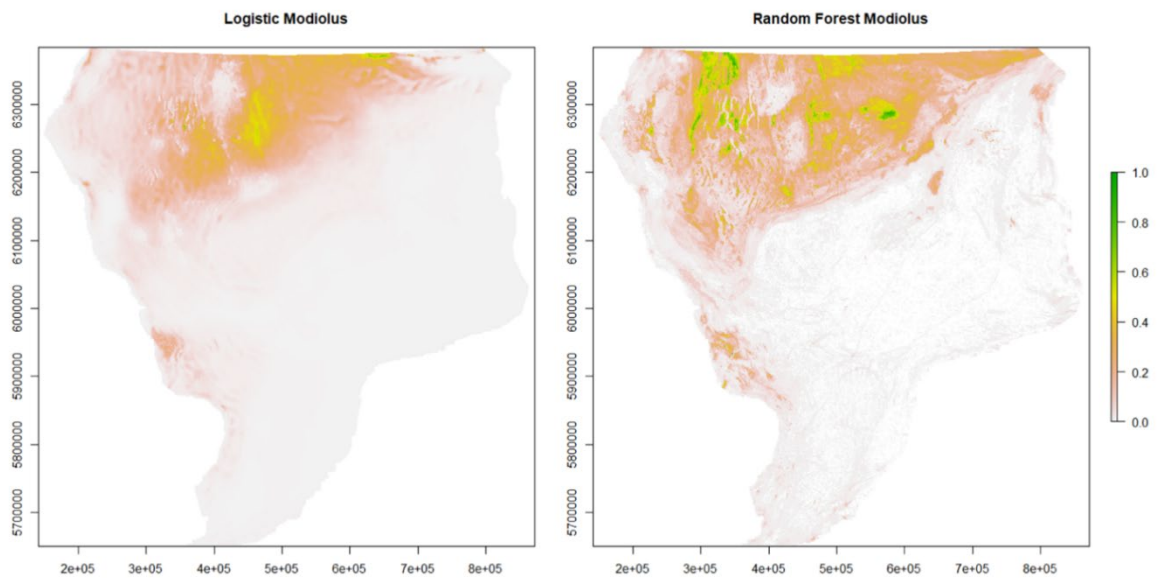


Figure 3.3. Results of the logistic (left) and random forest (right) regression models for *Modiolus modiolus*. The maps present the predicted probability of occurrence of the species over the entire North Sea.

3.3 Lanice conchilega

Lanice conchilega is one of the most frequently found species in the Greater North Sea. In the EMODnet data base, it occupies rank 9 in the list of the most frequent species. This is reflected in the distribution maps and the regression models for the southern part of the North Sea (Figure 3.4). *Lanice* has a clear preference for shallow areas with a reasonably high bottom shear stress and some influence of waves. It can be found from the beach down to a few tens of meters, in sandy areas. At a relatively small scale, the distribution of the species seems to be influenced by small-scale patterns in bathymetry, where its occurrence is linked to the relief of ridges and hollows in between. The random forest regression model picks up these features and predicts a quite fine-grained distribution pattern across the North Sea.

The distribution of *Lanice* was not explained by the intensity of fisheries. Research in the Voordelta, off the S-W Dutch coast, has shown that, if anything, it is positively related to shrimp fisheries intensity. This is most probably because it shares environmental preferences with shrimps, and apparently is also a sign that it is not negatively influenced by fisheries activities.

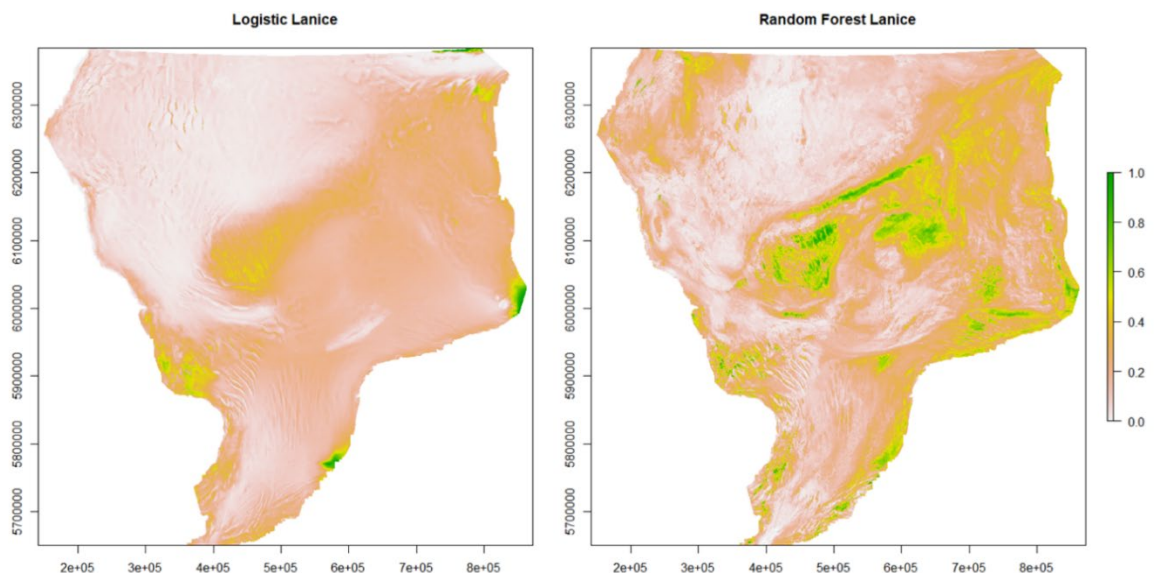


Figure 3.4. Results of the logistic (left) and random forest (right) regression models for *Lanice conchilega*. The maps present the predicted probability of occurrence of the species over the entire North Sea.

3.4 *Ostrea edulis*

The flat oyster is known to have been distributed widely in the North Sea until the end of the nineteenth century, when it was wiped out to a large degree by fisheries. Diseases, most prominently *Bonamia*, may have finished off this decline, so that the species has now almost disappeared from the North Sea. Recently, some populations that are resistant to *Bonamia* have developed in the estuaries in the S.W. Netherlands, and spread to the Voordelta, the harbour of Rotterdam and the Wadden Sea. Some attempts at reintroducing the species in the North Sea have taken place, and larvae have been detected in the waters of the North Sea even outside these introduction areas. The species is subject of many efforts for ecosystem restoration by reintroduction. The historical distribution of the species has been described by Bennema et al. (2020) for Dutch waters. Bennema et al. (2020) describe a general distribution area around the Oyster Grounds and eastward towards Helgoland. The 30m depth contour is described in historical records as a depth limit, suggesting the species was linked to areas with (at least) intermittent stratification in summer. The distribution patch is bounded to the east by a contour of decreasing salinity. Whereas lower salinity is not restrictive for the oyster, it may point to a boundary in currents, in particular the eastern boundary of the current that comes off the English coast and is directed towards the northern German Bight. This current is known to carry suspended solids, but likely also carries nutrients or phytoplankton from coastal enrichment.

In Belgian waters, a study by Houziaux et al. (2008) sheds light on the distribution in the early twentieth century. Here, the species was restricted to gravelly patches in between high sand banks.

We have not completed our data base with data from English waters, except for the expedition data provided by Bennema, that did not contain positive observations along the English coast. It is likely that the data we used are incomplete for this reason. The regression predictions should be fairly reliable for Dutch waters but may be incomplete in other parts of the North Sea. Within the scope of the present project, it was not possible to assemble a complete dataset for the historical distribution of flat oysters in the entire North Sea.

The regression models use sediment composition, bottom shear stress, depth and topography as elements to select the well-known 'oyster triangle' as its potential area of

development. Small patches may, in addition, develop in between sand banks. All areas have in common that the sediment is not composed of mobile sand but has a high proportion of either gravel or mud.

The historical distribution patterns are useful as a guidance for the spatial location of restoration efforts. They clearly show where efforts at reintroduction may lead to a population capable of self-reproduction. It is likely that substrate plays a key role in this process. Where hard substrate is provided, populations may develop also outside of the areas where they were present on natural substrate. The populations in Voordelta and Rotterdam harbour are examples of this possibility. We do not think that there are physiological limitations for flat oysters in the North Sea. Our study consequently cannot provide information on habitat suitability when artificial substrate is offered, but it seems likely that in that case most of the North Sea will allow oyster growth. With the present data, it cannot be decided without further experiments whether competition, predation and disease will allow the development of stable populations on all artificial substrates.

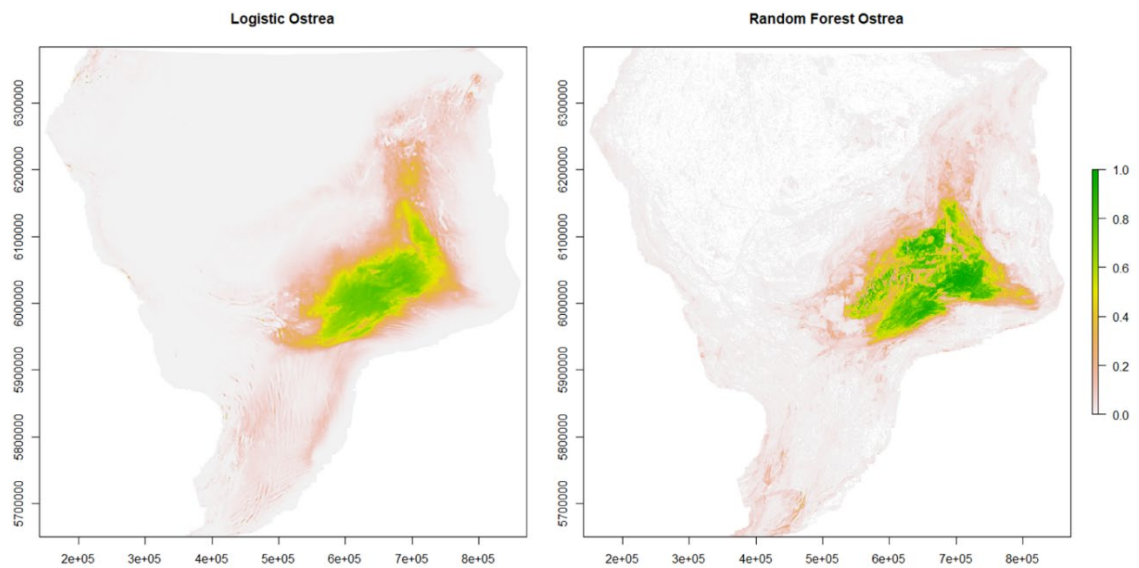


Figure 3.5. Results of the logistic (left) and random forest (right) regression models for Ostrea edulis. The maps present the predicted probability of occurrence of the species over the entire North Sea.

4 Acknowledgements

We are grateful to Oscar Bos (Wageningen Marine Research) for making the 'Datras' data available, EMODnet Biology for the availability of the collected data on macrobenthos in the North Sea, Karen van der Reijden for the environmental layers described in her paper, and Floris Bennema who provided his data files on oyster distribution.

5 References

Bennema, F.P., Engelhard, G.H., and Lindeboom, H. (2020). *Ostrea edulis* beds in the central North Sea: delineation, ecology, and restoration. ICES Journal of Marine Science 77(7-8), 2694-2705. doi: 10.1093/icesjms/fsaa134.

[Dataset] Herman, P.M.J., Stolte, W., and van der Heijden, L. (2020). Summary presence/absence maps of macro-endobenthos in the greater North Sea, based on nearly 100,000 samples from 65 assembled monitoring data sets. EMODNET Biology data product. Available: <https://www.emodnet-biology.eu/data-catalog?module=dataset&dasid=6617>.

Houziaux, J.-S., Kerckhof, F., Degrendele, K., Roche, M., and Norro, A. (2008). "The Hinder banks: yet an important area for the Belgian marine biodiversity ?", (ed.) B.S. Policy. (Brussels).

[Dataset] Stephens, D. (2015). North Sea and UK shelf substrate composition predictions, with links to GeoTIFFs. . doi: [doi:10.1594/PANGAEA.845468](https://doi.org/10.1594/PANGAEA.845468).

van der Reijden, K.J., Hintzen, N.T., Govers, L.L., Rijnsdorp, A.D., and Olf, H. (2018). North Sea demersal fisheries prefer specific benthic habitats. PLOS ONE 13(12), e0208338. doi: 10.1371/journal.pone.0208338.

van der Reijden, K. J., L. Koop, S. O'Flynn, S. Garcia, O. Bos, C. van Sluis, D. J. Maaholm, P. M. J. Herman, D. G. Simons, H. Olf, T. Ysebaert, M. Snellen, L. L. Govers, A. D. Rijnsdorp, and R. Aguilar. 2019. Discovery of Sabellaria spinulosa reefs in an intensively fished area of the Dutch Continental Shelf, North Sea. Journal of Sea Research 144:85-94.

A Appendices

A.1 R code for the analysis

A.1.1

Code chunk#1

```
# required packages
require(raster)
require(rgdal)
require(tidyverse)
require(tidync)
require(FNN)
require(randomForest)
# working directories
dcsm_dir <- paste0("p:/1204257-dcsmzuno/2013-2017/3D-DCSM-FM/A25_ntsu1/",
                  "DFM_OUTPUT_DCSM-FM_0_5nm")
emodnet_data_dir <- "./base_data/EMODnet_data"
DATRAS_data_dir <- paste0("./base_data/data_WMR/",
                          "2021-11-03-MONS-data-naar-Deltares/",
                          "DATRAS-Fish-surveys")
oyster_data_dir <- "./base_data/Ostrea"
coast_shape_dir <- "./Europe_coastline_shapefile"
stress_dir <- "./Environment/DCSM-FM/"
spec_data_dir <- "./Shapefiles_species"
WMR_spec_data_dir <- "./Shapefiles_species/WMR"
output_dir <- "./output"
# projection strings for spatial data
proWG<-CRS("+proj=longlat +datum=WGS84")
proUTM <- CRS("+proj=utm +zone=31 +ellps=GRS80 +units=m +no_defs")
```

A.1.2

Code chunk #2

```
if(! file.exists("./Environment/DCSM-FM/DCSM_mean_stress.tif")){
# read model output files and store mesh + bss in one data frame
for(filn in 0:19){
  filename<-paste0(dcsm_dir,"DCSM-FM_0_5nm_00",
                  formatC(filn, width = 2, format = "d", flag = "0"), "_fou.nc")
  tt <- tidync(filename) %>%
  activate("D3") %>%
  hyper_tibble() %>%
  select(mesh2d_face_x,
         mesh2d_face_y,
         mesh2d_flowelem_ba,
         mesh2d_flowelem_bl,
         mesh2d_flowelem_domain,
         mesh2d_flowelem_globalnr,
         mesh2d_fourier010_mean,
         mesh2d_fourier011_max)
  if(filn==0)tt2 <- tt else tt2 <- rbind(tt2,tt)
}
# reproject data in UTM
tt3<-tt2[,1:3]
coordinates(tt3) <- ~ mesh2d_face_x + mesh2d_face_y
projection(tt3) <- proWG
tt3<- spTransform(tt3,proUTM)
xymat <- as.matrix(coordinates(tt3))
tt2 <- tt2 %>%
```

```

mutate (x_utm=xymat[,1])%>%
mutate (y_utm=xymat[,2])
# define a regular raster to store the values with a 1 km resolution
r_mean_stress <- r_max_stress <- raster(ext=extent(-200000,900000,5300000,6600000),
                                         crs=proUTM,resolution=1000)
cor<-as.matrix(coordinates(r_mean_stress))
# for each raster cell, determine the nearest neighbour in model output and
# store this value in rasters
nn<-get.knnx(data=xymat,query=cor,k=1)
values(r_mean_stress)<-tt2$mesh2d_fourier010_mean[nn$nn.index]
values(r_max_stress)<-tt2$mesh2d_fourier011_max[nn$nn.index]
# use shapefile of European coastlines to blank out land
if(! file.exists(file.path(coast_shape_dir,"ecst.Rdata"))){
ecst<-readOGR("Europe_coastline_shapefile","Europe_coastline_poly")
ecst<-spTransform(ecst,proUTM)
ecst<-crop(ecst,extent(-500000,1000000,5000000,7000000))
save(ecst,file=file.path(coast_shape_dir,"ecst.Rdata"))
} else {
  load(file.path(coast_shape_dir,"ecst.Rdata"))
}
r_mean_stress <- mask(r_mean_stress,ecst,inverse=TRUE)
r_max_stress <- mask(r_max_stress ,ecst,inverse=TRUE)
raster::writeRaster(r_mean_stress,"./Environment/DCSM-FM/DCSM_mean_stress.tif",
                    overwrite=TRUE)
raster::writeRaster(r_max_stress,"./Environment/DCSM-FM/DCSM_max_stress.tif",
                    overwrite=TRUE)
}

# list of environmental factors
envies <- data.frame(name = c(
  "mean_stress",
  "depth",
  "bpi5",
  "bpi10",
  "bpi75",
  "wt_BPI",
  "meantemp",
  "difftemp",
  "salinity",
  "sand",
  "gravel",
  "mud",
  "beam_plaice",
  "beam_sole",
  "otter_mix"),
                    file = c(
  "./Environment/DCSM-FM/DCSM_mean_stress.tif",
  "./Environment/Environmental_factors/depth.tif",
  "./Environment/Environmental_factors/bpi5.tif",
  "./Environment/Environmental_factors/bpi10.tif",
  "./Environment/Environmental_factors/bpi75.tif",
  "./Environment/Environmental_factors/summed_weighted_BPI_SA.tif",
  "./Environment/Environmental_factors/meantemp.tif",
  "./Environment/Environmental_factors/difftemp.tif",
  "./Environment/Environmental_factors/salinity.tif",
  "./Environment/Environmental_factors/sand.tif",
  "./Environment/Environmental_factors/gravel.tif",
  "./Environment/Environmental_factors/mud.tif",
  "./Environment/Fisheries_data/Average_FI_Beam-Plaice.tif",
  "./Environment/Fisheries_data/Average_FI_Beam-Sole.tif",

```

```

"/Environment/Fisheries_data/Average_FI_Otter-Mix.tif"),
lowlim=c(
  0.1, # mean_stress
  -50, # depth
  -20, # bpi5
  -15, # bpi10
  -30, # bpi75
  6, # wt_BPI
  9.5, # meantemp
  9, # difftemp
  26, # salinity
  0.6, # sand
  0.02, # gravel
  0.02, # mud
  0.1, # beam_plaice
  0.1, # beam_sole
  0.1 # otter_mix
),
uplim=c(
  3, # mean_stress
  0, # depth
  20, # bpi5
  15, # bpi10
  30, # bpi75
  18, # wt_BPI
  13, # meantemp
  16, # difftemp
  34, # salinity
  0.99, # sand
  0.6, # gravel
  0.3, # mud
  2.1, # beam_plaice
  2.1, # beam_sole
  2.1 # otter_mix
)
)
)
# store rasters with env info if not yet done
rbas <- raster(envies$file[9])
for (i in 1:nrow(envies)){
  rfn<-paste0("./Environment/rasters/",envies$name[i],".grd")
  if(! file.exists(rfn)){
    r <- raster(envies$file[i])
    r2 <- resample(r,rbas)
    names(r2) <- envies$name[i]
    writeRaster(r2,file=rfn)
  }else{
    r2<-raster(rfn)
  }
  if(i==1)b<-brick(r2) else b <- addLayer(b,r2)
}
# prepare the data frame newdats, containing all environmental information for
# each node of the raster and used to make predictions
if(! file.exists("./Environment/newdats.Rdata")){
  newdats<-data.frame(mean_stress=values(raster::subset(b,"mean_stress")),
    depth=values(raster::subset(b,"depth")),
    bpi5=values(raster::subset(b,"bpi5")),
    bpi10 = values(raster::subset(b,"bpi10")),
    bpi75 = values(raster::subset(b,"bpi75")),
    meantemp=values(raster::subset(b,"meantemp")),

```



```

difftemp=values(raster::subset(b,"difftemp")),
salinity=values(raster::subset(b,"salinity")),
sand = values(raster::subset(b,"sand")),
gravel=values(raster::subset(b,"gravel")),
mud=values(raster::subset(b,"mud")),
lon=coordinates(b)[,1],
lat=coordinates(b)[,2]
newdats <- newdats %>%
  filter(!is.na(mean_stress))%>%
  filter(!is.na(depth)) %>%
  filter(!is.na(mud))
save(newdats,file="./Environment/newdats.Rdata")
} else {
load("./Environment/newdats.Rdata")
}

```

A.1.3 Code chunk #3

```

# load binary data with species presencs/absence data
load(file.path(emodnet_data_dir,"spe.Rdata"))
# function to extract a particular species from the data file and write as shapefile
extr_spec<-function(AphiaID,filnam){
col<-which(names(spe)==paste0("pa",AphiaID))
sabs<- cbind(spe[,1:4],spe[,col])
names(sabs)<-c("eventNumber","eventDate","lon","lat","presabs")
sabs<-sabs[!is.na(sabs$presabs),]
sabs$presabs<-ifelse(sabs$presabs,1,0)
coordinates(sabs)<- ~lon+lat
projection(sabs)<-proWG
writeOGR(sabs, file.path(spec_data_dir,filnam), filnam,
  driver="ESRI Shapefile",overwrite_layer = TRUE)
}
extr_spec(130867,"Sabellaria")
extr_spec(140467,"Modiolus")
extr_spec(131495,"Lanice")

```

A.1.4 Code chunk #4

```

# open file with DATRAS data
datras<-read.csv(file.path(DATRAS_data_dir,"biogene_rifsoorten_frisbe.csv"))
smeth<-unique(datras$TOR_CODE)
# only use data obtained with methods that can detect these four species
smeth<-smeth[c(grep("Boomkor",smeth),grep("GOV",smeth),grep("Otter",smeth))]
datras <- datras %>% filter(TOR_CODE %in% smeth)
# make list of samples
dat_samps <- datras %>%
  select(year,PGM_CODE,month,day,CODE,sample,latitude_s,longitude_s,
    DURATION,TOR_CODE) %>%
  distinct() %>%
  mutate(sampID=row_number())
# add unique sample number to all records in datras -> datcompl
datcompl<-datras %>%
  left_join (dat_samps,by=c("year","PGM_CODE","month","day","CODE","sample",
    "latitude_s","longitude_s","DURATION","TOR_CODE"))
# find the samples containing each of the four species
dat_mod_pos <- datcompl %>%
  filter(SCIENTIFIC_NAME=="*Modiolus modiolus*")
dat_sab_pos <- datcompl %>%
  filter(SCIENTIFIC_NAME=="*Sabellaria")
dat_lan_pos <- datcompl %>%
  filter(SCIENTIFIC_NAME=="*Lanice conchilega*")

```

```

dat_ost_pos <- datcompl %>%
  filter(SCIENTIFIC_NAME=="*Ostrea edulis*")
# add columns to the samples file indicating presence/absence of each of the species
dat_mod<- dat_samps %>%
  mutate(Modiolus=ifelse(sampID %in% dat_mod_pos$sampID,1,0),
         Sabellaria = ifelse(sampID %in% dat_sab_pos$sampID,1,0),
         Lanice = ifelse(sampID %in% dat_lan_pos$sampID,1,0),
         Ostrea = ifelse(sampID %in% dat_ost_pos$sampID,1,0)) %>%
  filter(! is.na(longitude_s) & ! is.na(latitude_s)) %>%
  filter(latitude_s > 49) %>%
  mutate(eventNummer = sampID+100000,
         eventDate = as.Date(paste(year, month, day,sep="-"), "%Y-%m-%d"),
         decimalLongitude=longitude_s,
         decimalLatitude = latitude_s) %>%
  select(eventNummer,eventDate,Modiolus,Sabellaria,Lanice,Ostrea,
         decimalLongitude,decimalLatitude)
# save file
save(dat_mod,file=file.path(DATRAS_data_dir,"dat_mod.Rdata"))
# transform into spatial object
coordinates(dat_mod) <- ~ decimalLongitude + decimalLatitude
projection(dat_mod) <- proWG
# save shape files with the species observations
tt<- dat_mod[,"Modiolus"]
names(tt)<-"presabs"
writeOGR(tt, file.path(WMR_spec_data_dir,"Datras_Modiolus"), "Datras_Modiolus",
         driver="ESRI Shapefile",overwrite_layer = TRUE)
tt<- dat_mod[,"Lanice"]
names(tt)<-"presabs"
writeOGR(tt, file.path(WMR_spec_data_dir,"Datras_Lanice"), "Datras_Lanice",
         driver="ESRI Shapefile",overwrite_layer = TRUE)
tt<- dat_mod[,"Ostrea"]
names(tt)<-"presabs"
writeOGR(tt, file.path(WMR_spec_data_dir,"Datras_Ostrea"), "Datras_Ostrea",
         driver="ESRI Shapefile",overwrite_layer = TRUE)
tt<- dat_mod[,"Sabellaria"]
names(tt)<-"presabs"
writeOGR(tt, file.path(WMR_spec_data_dir,"Datras_Sabellaria"), "Datras_Sabellaria",
         driver="ESRI Shapefile",overwrite_layer = TRUE)

```

A.1.5 Code chunk #5

```

# 1. Generate pseudo-absences and pseudo-presences in the mapped area of Bennema
tt<-readOGR(file.path(oyster_data_dir,"old_maps.shp"))
tt<-spTransform(tt,proUTM)
tt$id<-1
r<-subset(b,"mean_stress")
extb<-extent(b)
# generate random points
ngp<-0
rp<-data.frame(x=NA,y=NA,oyster=NA)
while(ngp<3000){
  rpt <- data.frame(x=extb[1]+(extb[2]-extb[1])*runif(1000),
                  y=extb[3]+(extb[4]-extb[3])*runif(1000),oyster=NA)
  rrp<-raster::extract(r,rpt[,1:2])
  rrp<-which(!is.na(rrp))
  if(ngp==0)rp<-rpt[rrp,] else rp <- rbind(rp,rpt[rrp,])
  ngp<-ngp+length(rrp)
}

```

```

# make spatial to get projection right
rps <- rp
coordinates(rps) <- ~x+y
projection(rps)<-proUTM
rps$oyster <- ifelse(is.na(over(rps,tt)),0,1) *
  as.numeric(runif(length(rps$oyster))>0.3)
rps<-spTransform(rps,proWG)
# and store as data frame
rp<-as.data.frame(rps)
rp <- rp %>%
  mutate(eventNummer=row_number()+300000,
    eventDate=as.Date("1880-01-01",format="%Y-%m-%d"),
    decimalLongitude=x,
    decimalLatitude=y,
    Ostrea=oyster,
    Sabellaria=NA,
    Modiolus=NA,
    Lanice=NA) %>%
  select(decimalLongitude,decimalLatitude,eventNummer,eventDate,
    Sabellaria,Lanice,Modiolus,Ostrea)
save(rp,file=file.path(oyster_data_dir,"rp.Rdata"))

# 2. Retrieve oyster information from expeditions and Gilson
ostrea_HW <- read.csv(file.path(oyster_data_dir,"Huxley_Wodan_ostrea.csv"),
  fileEncoding = 'UTF-8-BOM')
ostrea_P <- read.csv(file.path(oyster_data_dir,"Poseidon_ostrea.csv"),
  fileEncoding = 'UTF-8-BOM')
ost<-rbind(ostrea_HW,ostrea_P)
ost <- ost[!(ost$lat==0 & ost$lon==0),]
ost$presabs <- ost$ostrea_edulis_pres
ost <- ost[,c(6,7,21)]
coordinates(ost)<- ~lon+lat
projection(ost)<-proWG
gils<-readOGR(dsn=file.path(oyster_data_dir,"points_Gilson"),
  layer="points_Gilson")
gils<-gils[,-1]
ost<- rbind(ost,gils)
# write Ostrea shapefile
writeOGR(ost, file.path(spec_data_dir,"Ostrea"), "Ostrea",
  driver="ESRI Shapefile",overwrite_layer = TRUE)
# reconstruct data frame with all observations, restructure, and store as binary file
ost_df<-data.frame(decimalLongitude=coordinates(ost)[,1],
  decimalLatitude=coordinates(ost)[,2],
  Ostrea=ost$presabs,
  Lanice=NA,
  Modiolus=NA,
  Sabellaria=NA,
  eventNummer=NA,
  eventDate=NA)
ost_df <- ost_df %>% mutate(eventNummer=row_number()+200000)
save(ost_df,file=file.path(oyster_data_dir,"ost_df.Rdata"))

```

A.1.6

Code chunk #6

```

if(! file.exists(file.path(emodnet_data_dir,"specenv.Rdata"))){
  # read species distribution from EMODnet data
  load(file.path(emodnet_data_dir,"spe.Rdata"))
  specenv <- spe %>%
    select(eventNummer,eventDate,decimalLongitude,decimalLatitude,
           pa130867,pa140467,pa131495) %>%
    mutate(Sabellaria = ifelse(pa130867,1,0),
           Modiolus = ifelse(pa140467,1,0),
           Lanice = ifelse(pa131495,1,0),
           Ostrea = NA) %>%
    select(-pa130867,-pa140467,-pa131495)
  # add DATRAS information for Sabellaria and Modiolus
  load(file.path(DATRAS_data_dir,"dat_mod.Rdata"))
  dat_mod <- dat_mod %>%
    mutate(Lanice = NA, Ostrea = NA)
  specenv<-rbind(specenv,dat_mod)
  # add historical data for Ostrea
  load(file.path(oyster_data_dir,"ost_df.Rdata"))
  specenv<-rbind(specenv,ost_df)
  # add historical data oyster based on maps (pseudo data points)
  load(file.path(oyster_data_dir,"rp.Rdata"))
  specenv<-rbind(specenv,rp)
  # make spatial
  coordinates(specenv)<- ~ decimalLongitude + decimalLatitude
  projection(specenv) <- proWG
  specenv <- spTransform(specenv,proUTM)
  # add environmental information to species distribution data
  for (i in 1:nrow(envies)){
    r <- subset(b,envies$name[i])
    specenv$newenv <- raster::extract(r,specenv)
    names(specenv)[which(names(specenv)=="newenv")]<-envies$name[i]
  }
  # store binary file
  save(specenv,file=file.path(emodnet_data_dir,"specenv.Rdata"))
}else{
  load(file.path(emodnet_data_dir,"specenv.Rdata"))
}

```

A.1.7

Code chunk #7

```

# now make the plots
specmaxs<-c(0.5,0.07,0.4,0.5)
specnams<-c("Sabellaria","Modiolus","Lanice","Ostrea")
nspec<-length(specnams)
for(spec in 1:nspec){
  specnam <- specnams[spec]
  specenvi<-as.data.frame(specenv)
  colspec <- which(names(specenvi) == specnam)
  specenvi<-specenvi[!is.na(specenvi[,colspec]),]
  specenvi$e_e <- specenvi[,colspec]
  spmaxy <- specmaxs[spec]
  for(env in 1:nrow(envies)){
    env_name <- envies$name[env]
    colenv <- which(names(specenvi)==env_name)
    specenvi$e_e <- specenvi[,colenv]
    emin <- envies$lowlim[env]
    emax <- envies$uplim[env]
    specenvi <- specenvi %>% drop_na(e_e)
    specenvi$qg <- cut(specenvi$e_e,

```

```

        breaks=unique(c(min(specenvi$e_e-1),
                        envies$lowlim[env],
                        quantile(specenvi$e_e[specenvi$e_e>envies$lowlim[env]
                                &specenvi$e_e<=envies$uplim[env]],
                                probs=seq(0.1, 1, 0.1)),
                        max(specenvi$e_e+1))),labels=F)
summ_ntile <- specenvi %>%
  mutate(qg = ntile(e_e, 12)) %>%
  group_by(qg) %>%
  summarise (meanenv = mean(e_e),
            minenv = min(e_e),
            maxenv = max(e_e),
            meanspec=mean(s_e))

plot(summ_ntile$meanenv,summ_ntile$meanspec,main=paste(specnam,"vs",env_name),
      xlab=paste(env_name),ylab="Class mean occurrence",
      ylim=c(0,spmaxy),xlim=c(min(summ_ntile$minenv),max(summ_ntile$maxenv)))
arrows(summ_ntile$minenv,summ_ntile$meanspec,summ_ntile$maxenv,summ_ntile$mea
nspec,code=0)
}
}

```

A.1.8

Code chunk #7

first clear memory

```

rm(b,dat_mod,dat_mod_pos,dat_ost_pos,dat_sab_pos,dat_lan_pos,dat_samps,datcompl,
  datras,gils,ost,ost_df,ostrea_HW,ostrea_P,r,r2,rbas,rp,rps,rpt,spe,
  summ_ntile,tablel,tt)
void<-as.vector(rep(" ",21))
outtbl<-data.frame(terms=void,coefSabel=void,PrSabel=void,coefModiol=void,
  PrModiol=void,coefLanic=void,prLanic=void,coefOstrea=void,
  prOstrea=void)

```

logistic regression

```

for(spec in 1:nspec){
  specnam<-specnams[spec]
  specenvi<-as.data.frame(specenv)
  colspec <- which(names(specenvi) == specnam)
  specenvi<-specenvi[!is.na(specenvi[,colspec]),]
  specenvi$s_e <- specenvi[,colspec]
  specenvi <- specenvi %>% select(- wt_BPI)
  specenvi$dum <- apply(specenvi[,7:20],1,sum)
  specenvi <- specenvi[!is.na(specenvi$dum),]
  ggg <- glm(s_e ~ mean_stress + l(mean_stress^2)+
    depth + l(depth^2) +
    bpi5 + l(bpi5^2) +
    bpi10 + l(bpi10^2) +
    bpi75 + l(bpi75^2) +
    meantemp + l(meantemp^2)+
    difftemp + l(difftemp^2)+
    salinity + l(salinity^2)+
    gravel + l(gravel^2) +
    mud + l(mud^2),
    specenvi,family="binomial")
  sggg<-summary(ggg)
  save(sggg,file=file.path(output_dir,paste0("logit_model_",specnam,".Rdata")))
  outtbl[,1]<-row.names(sggg$coefficients)
  outtbl[,spec*2]<-sggg$coefficients[,1]
}

```

```

tt<-sggg$coefficients[,4]
outtble[,spec*2+1] <- ifelse(tt<0.10&tt>0.05, ".",
                             ifelse(tt<0.05&tt>0.01, "**",
                                     ifelse(tt<0.01&tt>0.001, "***",
                                             ifelse(tt<0.001, "****", ""))))
newdats$preds_e <- predict(ggg,newdata=newdats,type="response")
predrast <- rasterFromXYZ(newdats[,c("lon","lat","preds_e")],crs=proUTM)
plot(predrast,main=paste("Predicted presence of ",specnam))
raster::writeRaster(predrast,
                    file=file.path(output_dir,
                                    paste0("logit_pred_",specnam,".tif")),
                    overwrite=TRUE)
}
save(outtble,file=file.path(output_dir,"outtble.Rdata"))

```

A.1.9 Code chunk #9

```

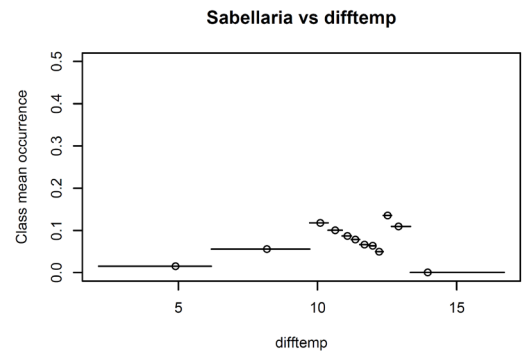
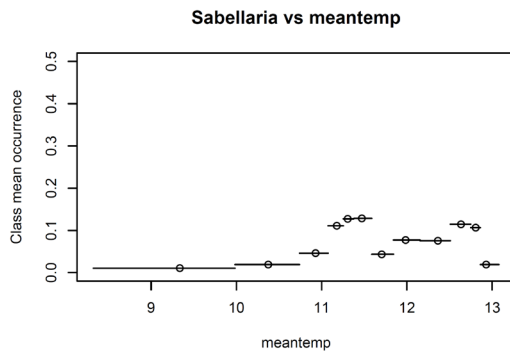
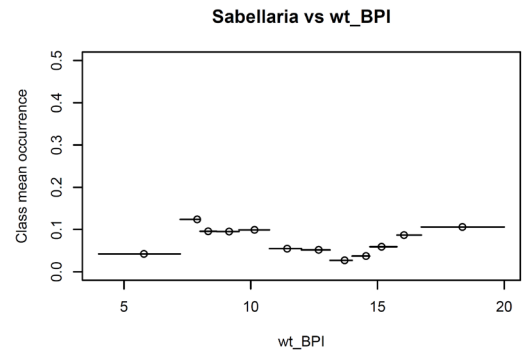
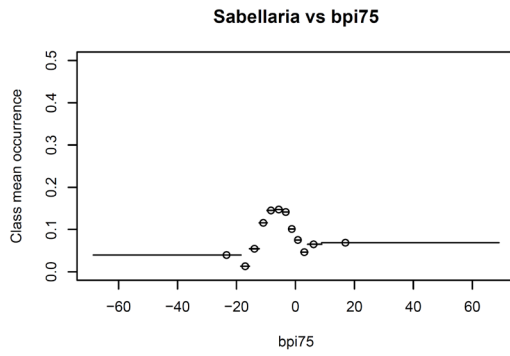
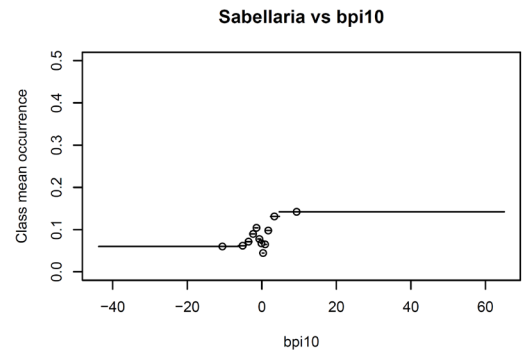
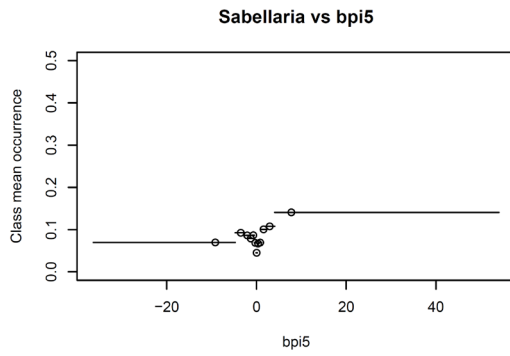
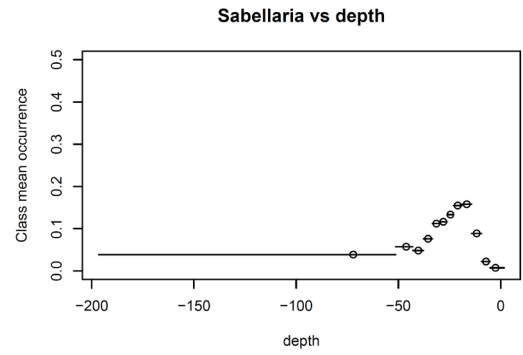
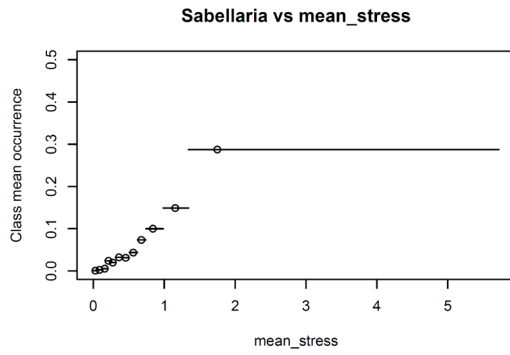
void<-c("mean_stress","depth","bpi5","bpi10","bpi75","meantemp","difftemp",
        "salinity","gravel","mud")
outtblerf<-data.frame(terms=void,IM_Sabel=void,INP_Sabel=void,IM_Modiol=void,
                      INP_Modiol=void,IM_Lanice=void,Inp_Lanice=void,IM_Ostrea=void,
                      INP_Ostrea=void)
for(spec in 1:nspec){
  specnam<-specnams[spec]
  specenvi<-as.data.frame(specenv)
  colspec <- which(names(specenvi) == specnam)
  specenvi<-specenvi[!is.na(specenvi[,colspec]),]
  specenvi$s_e <- specenvi[,colspec]
  specenvi <- specenvi %>% select(- wt_BPI)
  specenvi$dum <- apply(specenvi[,7:20],1,sum)
  specenvi <- specenvi[!is.na(specenvi$dum),]
  rf <- randomForest(s_e ~ mean_stress +
                    depth +
                    bpi5 +
                    bpi10 +
                    bpi75 +
                    meantemp +
                    difftemp +
                    salinity +
                    gravel +
                    mud,
                    specenvi,ntree=1000,importance=TRUE)

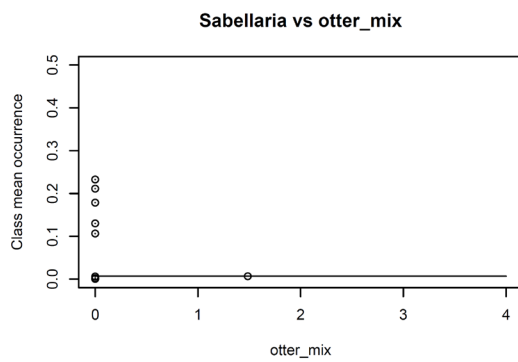
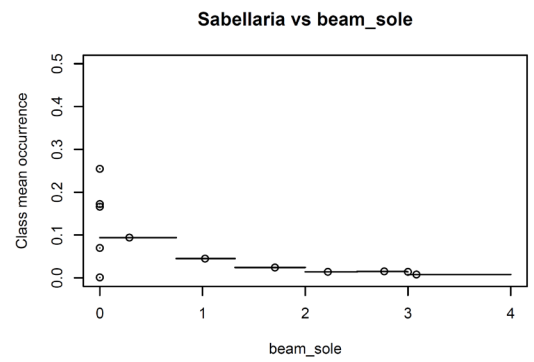
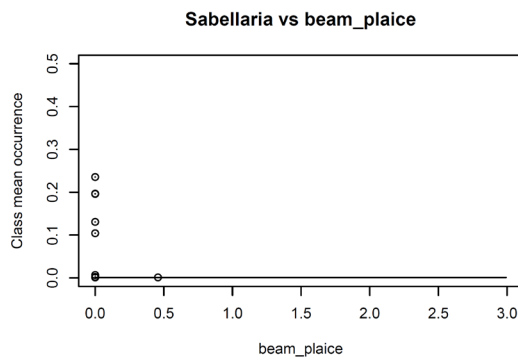
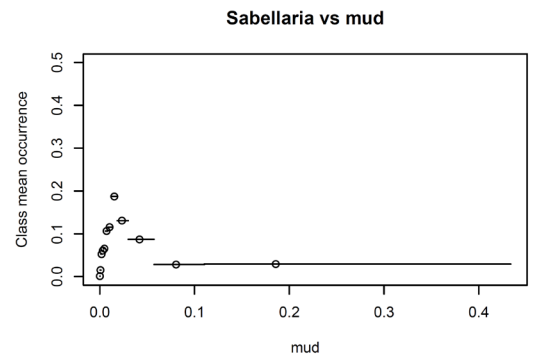
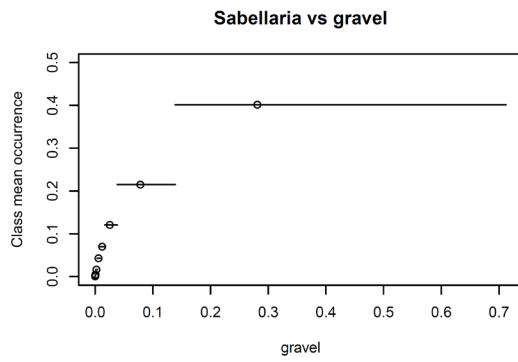
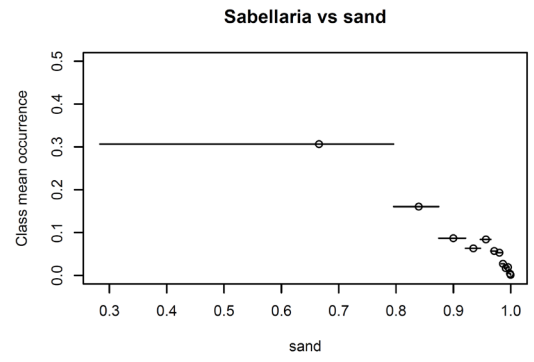
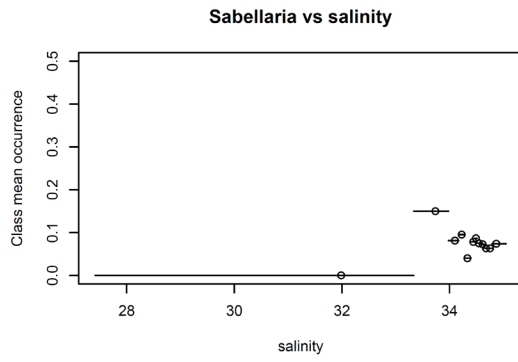
  print(rf)
  print(rf$importance)
  varImpPlot(rf)
  save(rf,file=file.path(output_dir,paste0("rf_model_",specnam,".Rdata")))
  outtblerf[(spec*2):(spec*2+1)]<-importance(rf)
  newdats$preds_e <- predict(rf,newdata=newdats)
  predrast <- rasterFromXYZ(newdats[,c("lon","lat","preds_e")],crs=proUTM)
  plot(predrast,main=paste("Predicted presence of ",specnam))
  raster::writeRaster(predrast,
                    file=file.path(output_dir,
                                    paste0("RF_pred_",specnam,".tif")),
                    overwrite=TRUE)
}
save(outtblerf,file="./output/outtblerf.Rdata")

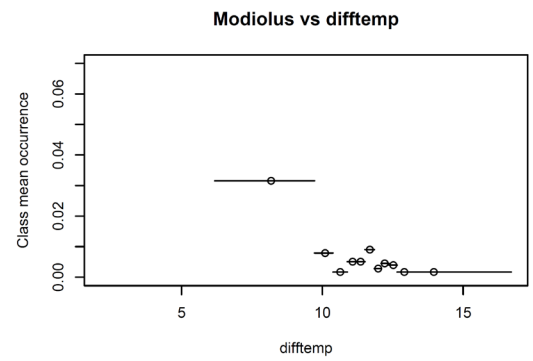
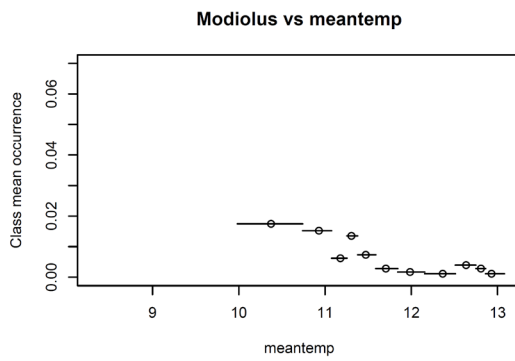
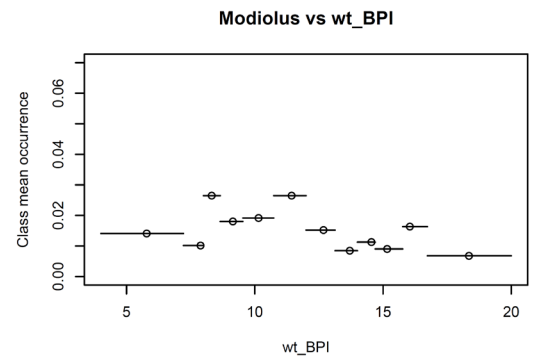
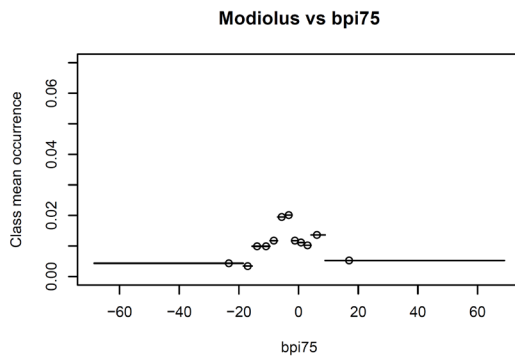
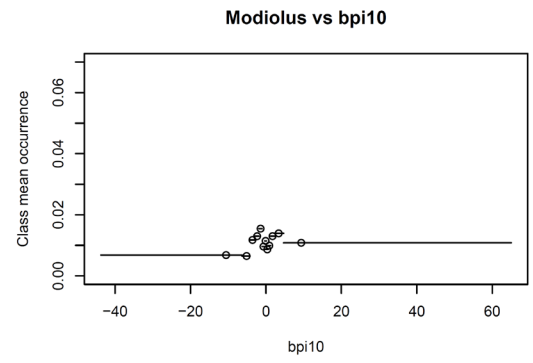
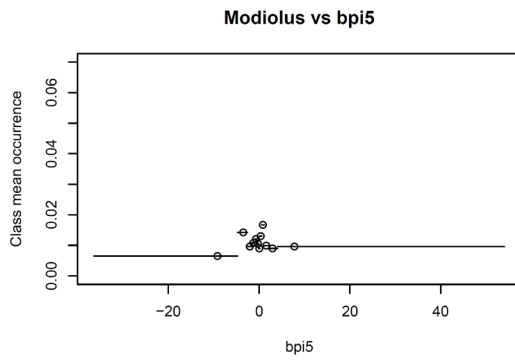
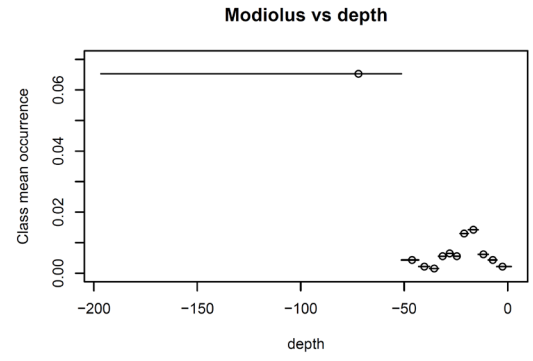
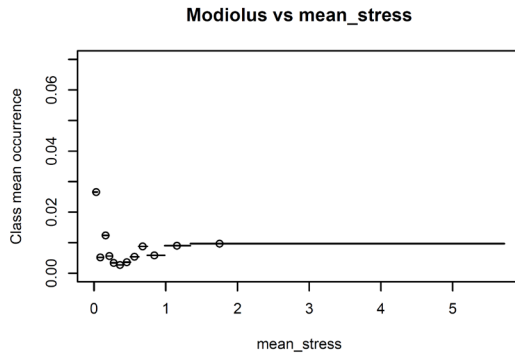
```

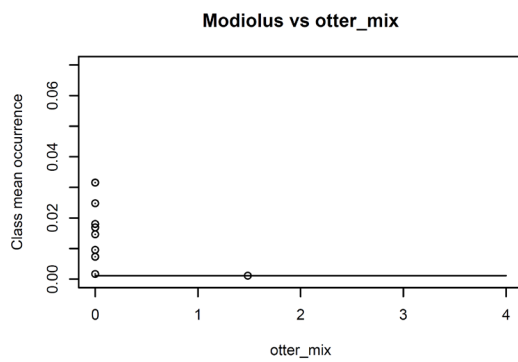
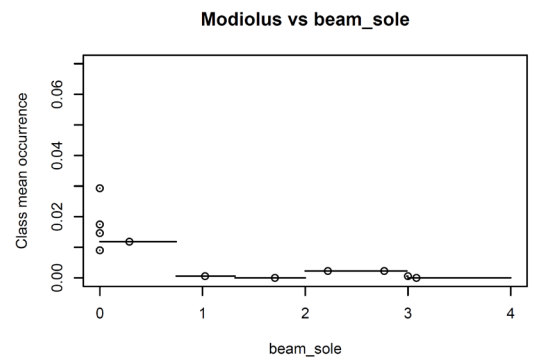
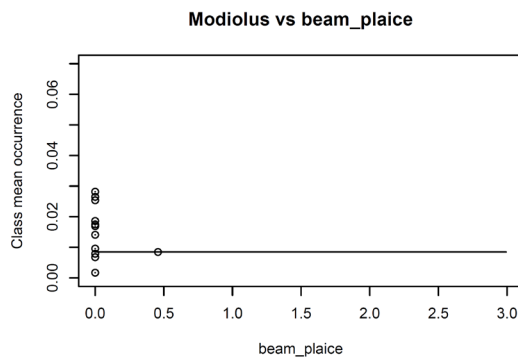
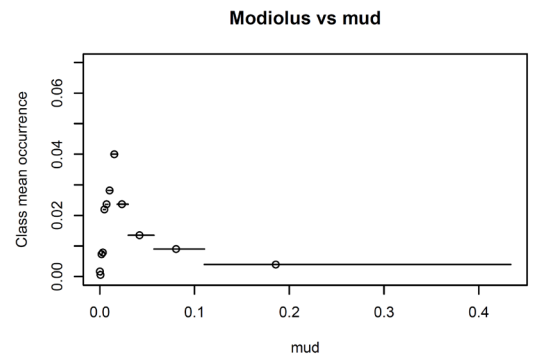
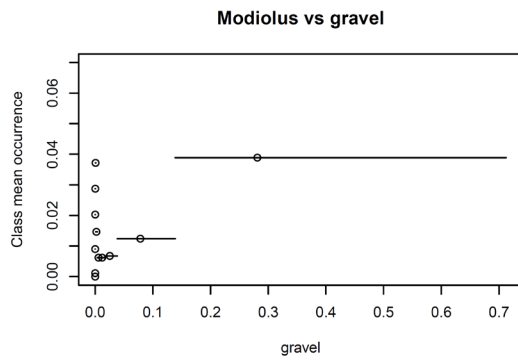
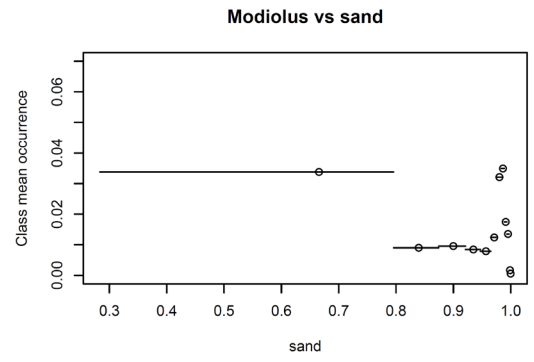
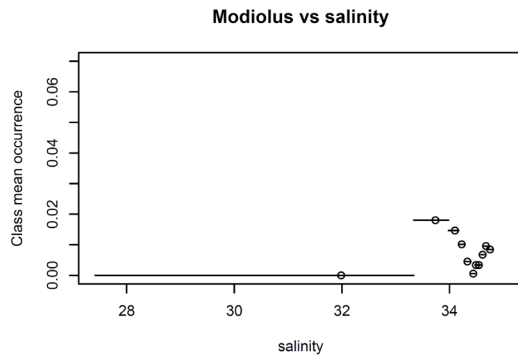
A.2 Exploratory species-environment plots

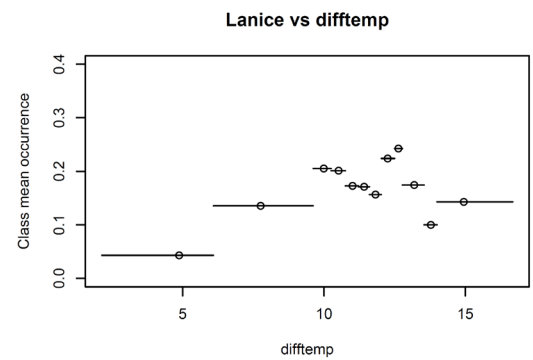
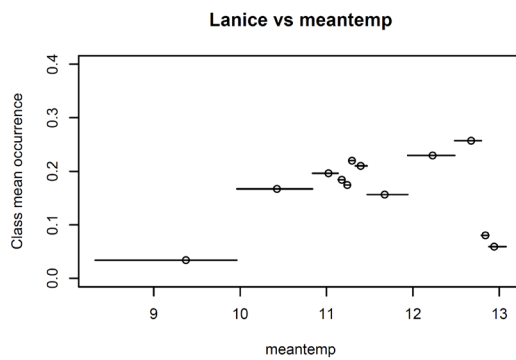
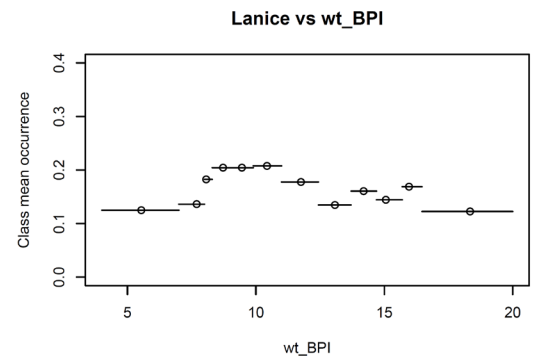
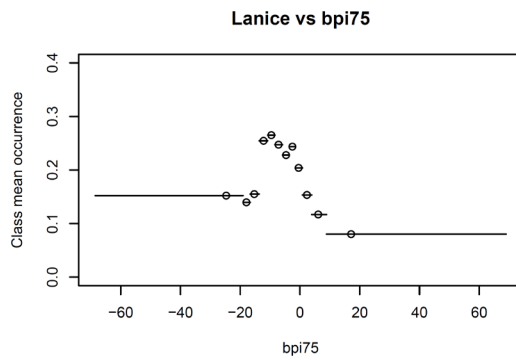
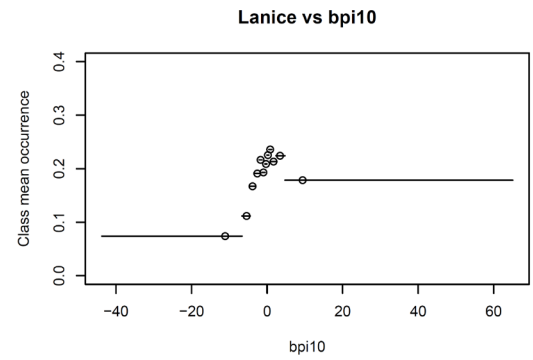
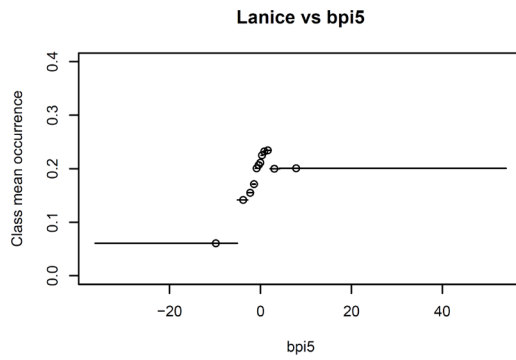
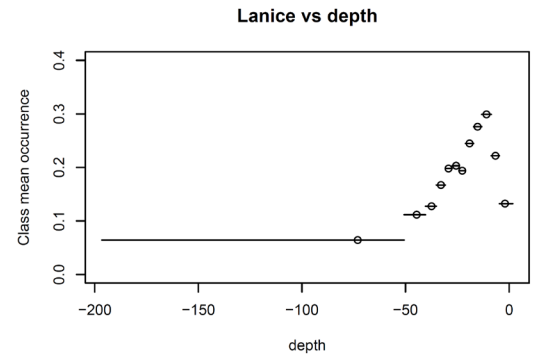
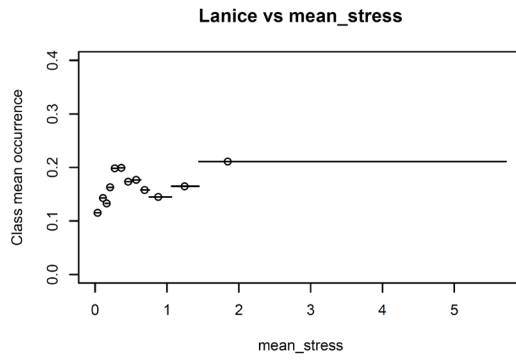
Plots of raw data of species occurrence versus environmental factors in the database. For each plot, the observations are split in twelve groups of increasing value of the environmental variable. Each group has an equal number of observations. Per group, the mean occurrence of the species in the group is plotted versus the mean value of the environmental variable in the group. Ranges of the environmental variable are also indicated. These plots are purely exploratory.

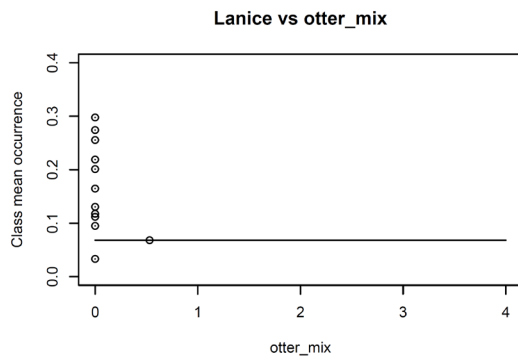
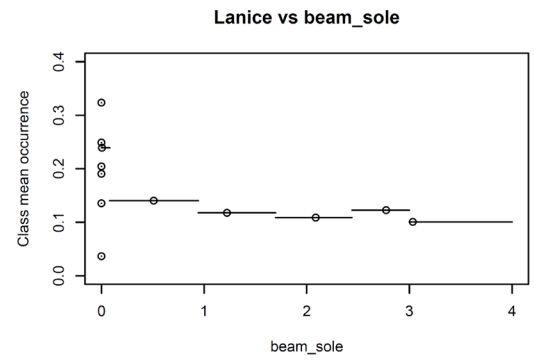
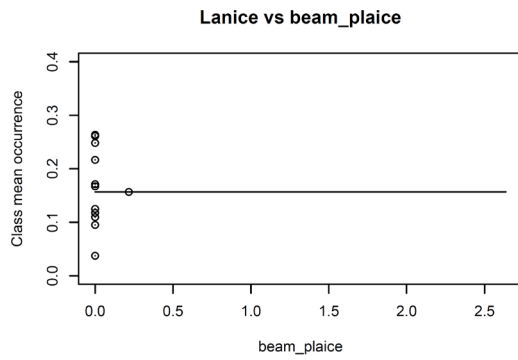
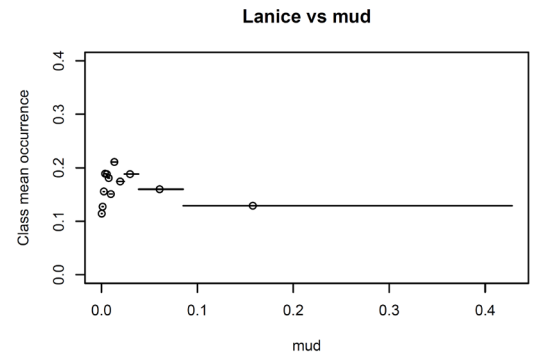
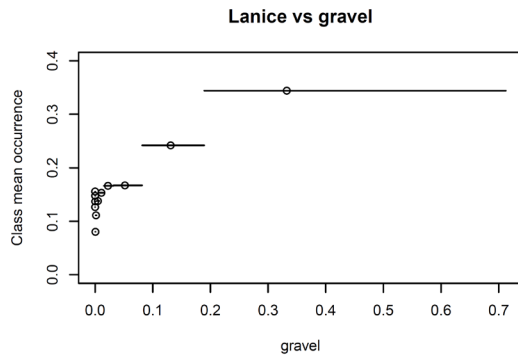
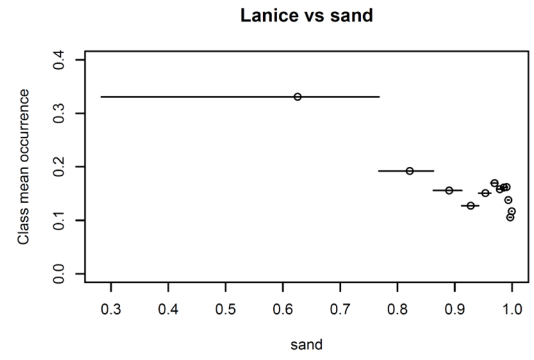
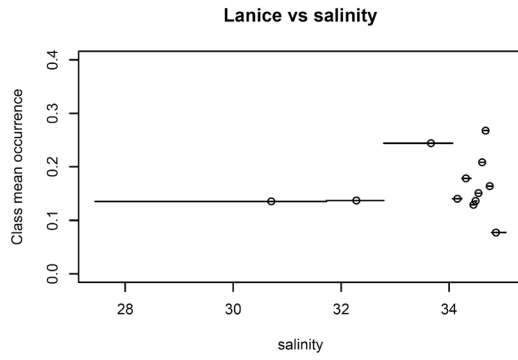


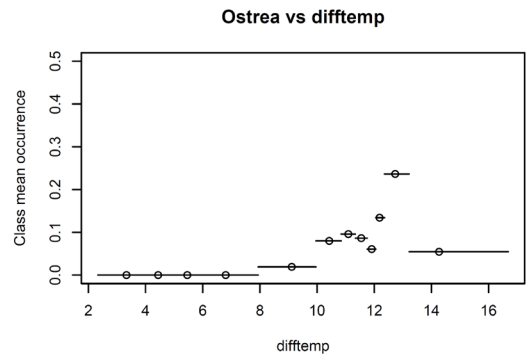
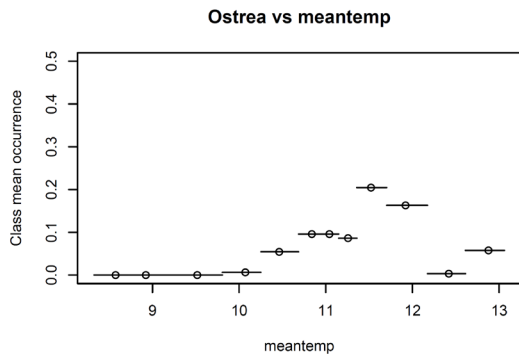
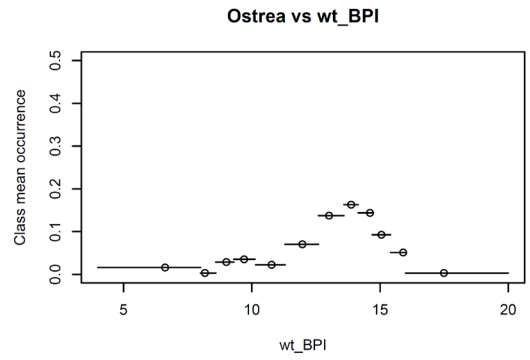
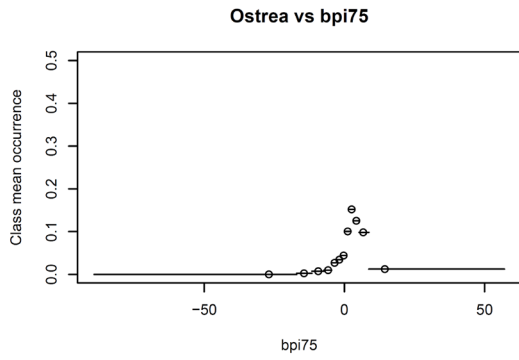
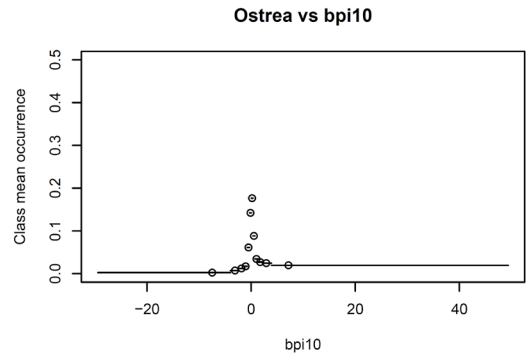
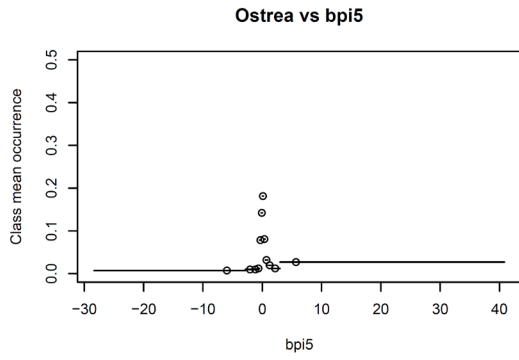
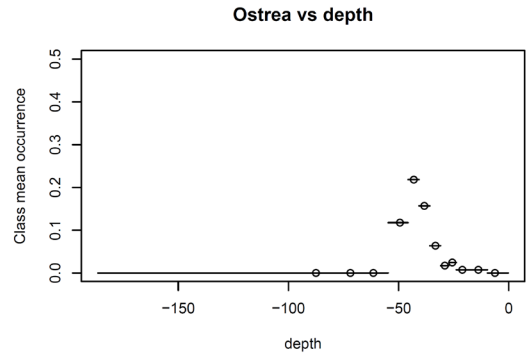
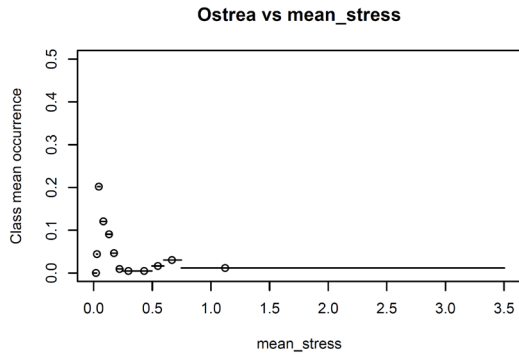


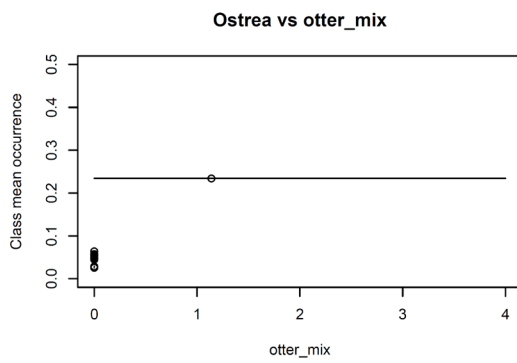
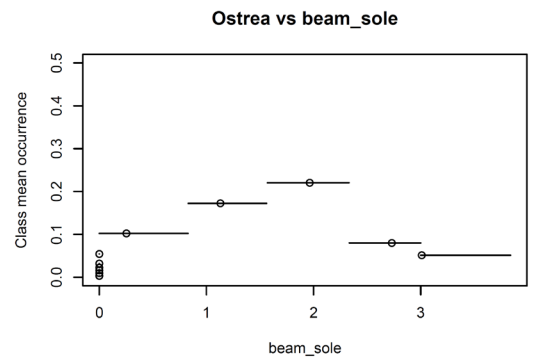
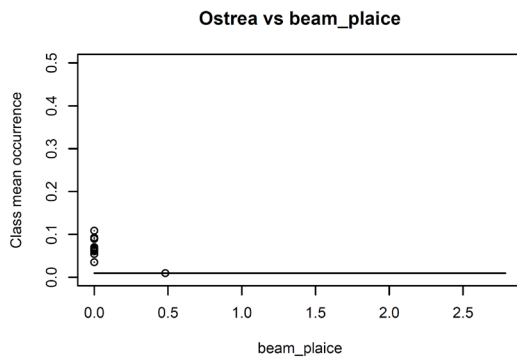
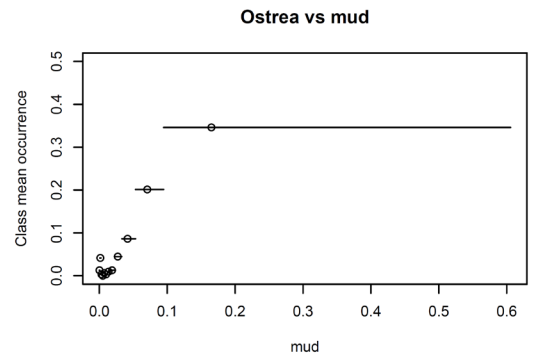
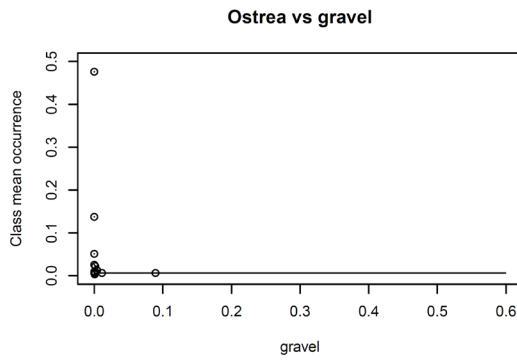
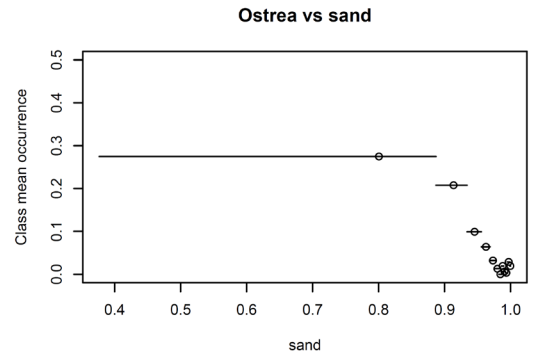
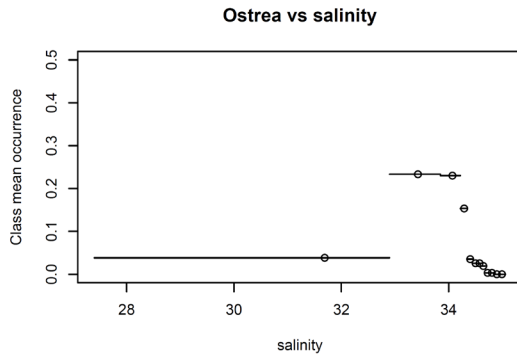












Deltares is an independent institute for applied research in the field of water and subsurface. Throughout the world, we work on smart solutions for people, environment and society.

Deltares

www.deltares.nl

Handtekening: 

E-mail: peter.herman@deltares.nl

Handtekening: 

E-mail: luca.vanduren@deltares.nl

Handtekening: 

E-mail: Paul.Saager@deltares.nl