CHAPTER 7

APPLICATION TO MACROBENTHIC SPECIES

CHAPTER 7

APPLICATION TO MACROBENTHIC SPECIES

Nature conservation involves considering many different aspects of the ecosystem (Villa *et al.*, 2002; Pomeroy *et al.*, 2004; Derous, 2008). The evaluation of the potential biological value of an area should be based on different biological components of the area (Derous, 2008). Here, we focused on the macrobenthos. Within this group we considered two potentially important species: *Lanice conchilega* and *Ensis directus*.

Lanice conchilega is an important ecosystem engineer which may entail high species richness when appearing in high densities (Rabaut *et al.*, 2007; Van Hoey *et al.*, 2008). Moreover, *L. conchilega* has the capacity to double the biodiversity of the *Abra alba* community (Van Hoey, 2006), a community which is characterised by both high macrobenthic densities and high species richness. Therefore, locations with dense aggregations of *L. conchilega* species have been suggested for nature conservation within the framework of the Habitats Directive (Degraer *et al.*, 2009). *Ensis directus*, on the other hand, is an invasive species which might compete for space and resources with the species rich *Abra alba* community. Therefore, estimating the potential distribution of this invasive species can indicate if an effect might be expected. Moreover, the models may contribute to evaluate the feasibility of a targeted *Ensis* fishery within Belgian waters.

Since these models are potentially being used for management purposes, they should be beyond discussion. Therefore, the techniques developed in Chapters 4 to 6 are applied here. For a detailed description of the techniques, we refer to the previous chapters and Addendum 1. The differences or additional calculations will be pointed out in the text.

LANICE CONCHILEGA (PALLAS, 1766) AGGREGATIONS

Introduction

A multi-criteria analysis tool as a decision tool for marine management, considering different aspects of marine ecosystems such as seabirds, macrobenthos, epibenthos, hyperbenthos and ecosystem processes, has been developed (Derous, 2008). Herein, *Lanice conchilega*, a member of the macrobenthos, has been suggested as a habitat forming keystone species (Derous, 2008). This species is considered to be important in the framework of nature conservation (Van Hoey, 2006; Godet *et al.*, 2008; Toupoint *et al.*, 2008; Rabaut *et al.*, 2009). Therefore, species distribution models of *L. conchilega* can be very useful to delineate areas of interest for nature conservation.

Lanice conchilega or sand mason is a polychaete, which builds linear tubes consisting of coarse sand grains cemented with mucus (Jones and Jago, 1993). The tube can reach a diameter of 5 mm and a length of 65 cm (Ziegelmeier, 1952). The tube is located mainly in the sediment, and only one to four centimetres protrude in the water column. This species has the ability to build dense aggregates and patches with more than 1500 ind.m^{-2} are not uncommon (Zühlke, 2001). From densities of around 500 ind.m^{-2} the tubes start consolidating the sediment and create a surface structure of gentle mounds ('reefs') (Rabaut *et al.*, 2009). The tubes compact the sediment and increase the rigidity of the sea floor (Jones and Jago, 1993). Moreover, these tubes trap sediment and change the hydrodynamics locally (Eckman, 1983). In this way the species can change the physical environment. In addition, it affects the biological community: the diversity of the surrounding benthic community increases with increasing densities of *L. conchilega*, and the diversity displays an optimum at around 1000 ind.m^{-2} (Rabaut *et al.*, 2007; Van Hoey *et al.*, 2008). Many aspects may contribute to the higher diversity: the lower flow current near the bottom attracts associated benthos, the movement of the polychaete in the tube may function as an oxygen pump (Braeckman *et al.*, 2010), and the biogenic structures are supposed to function as a shelter (Forster and Graf, 1995) and as feeding ground (Rabaut *et al.*, 2010). *Lanice* reefs attract flat fish, such as *Solea solea*, and may function as nursery grounds (Vanaverbeke *et al.*, 2009; Rabaut *et al.*, 2010).

The species is vulnerable to anthropogenic impacts such as sludge disposal (Witt *et al.*, 2004) and scraping of the sediment (Toupoint *et al.*, 2008). The reef structure can persist under intermediate beam-trawling pressure (Rabaut *et al.*, 2008); however the associated fauna is significantly impacted. Under intensive beam-trawling, the reef structure will eventually disappear (Rabaut, 2009).

Here, we focus on habitat suitability modelling of *L. conchilega*. Different thresholds of *L. conchilega* densities are considered since the potential for altering the habitat structure and enhancing the biodiversity of the surrounding community, are related with the density of the species (Rabaut *et al.*, 2007; Van Hoey *et al.*, 2008).

Material and methods

Research area

This work is done in the framework of evaluating areas for their potential use as protected areas in the Belgian Part of the North Sea (Degraer *et al.*, 2009). Therefore, the research area was restricted to the Belgian Part of the North Sea.

Lanice conchilega data

The *L. conchilega* data was retrieved from the MacroDat database (Degraer *et al.*, 2003a). This database is a compilation of macrobenthos data of the BPNS and beaches from 1163 stations (Fig. 7.1) taken within the time frame 1971-2008. From this database, the stations and densities with *L. conchilega* records were extracted, resulting in a dataset consisting of 231 stations where densities between 3 and 13 000 individuals per square meter were recorded. From a conservational point of view, especially the samples with high *Lanice* densities are of interest (Rabaut *et al.*, 2007; Van Hoey *et al.* 2008). Therefore, four different density thresholds were used: 1) at least 1 ind.m^{-2}, 2) at least 100 ind.m^{-2}, 3) at least 500 ind.m^{-2} and 4) aggregations with at least 1000 ind.m^{-2}. Most samples have densities between 3 and 80 individuals per sample, only a limited number of samples show high densities. The density data was then converted to presence data taking into consideration the density threshold values. In other words, if the species was observed in densities higher than the threshold, it was considered to be present. For the different densities respectively 231, 86, 42 and 29 stations were selected (Fig. 7.1). The models resulting from this data will be further referred to as Lanice1, Lanice100, Lanice500 and Lanice1000.

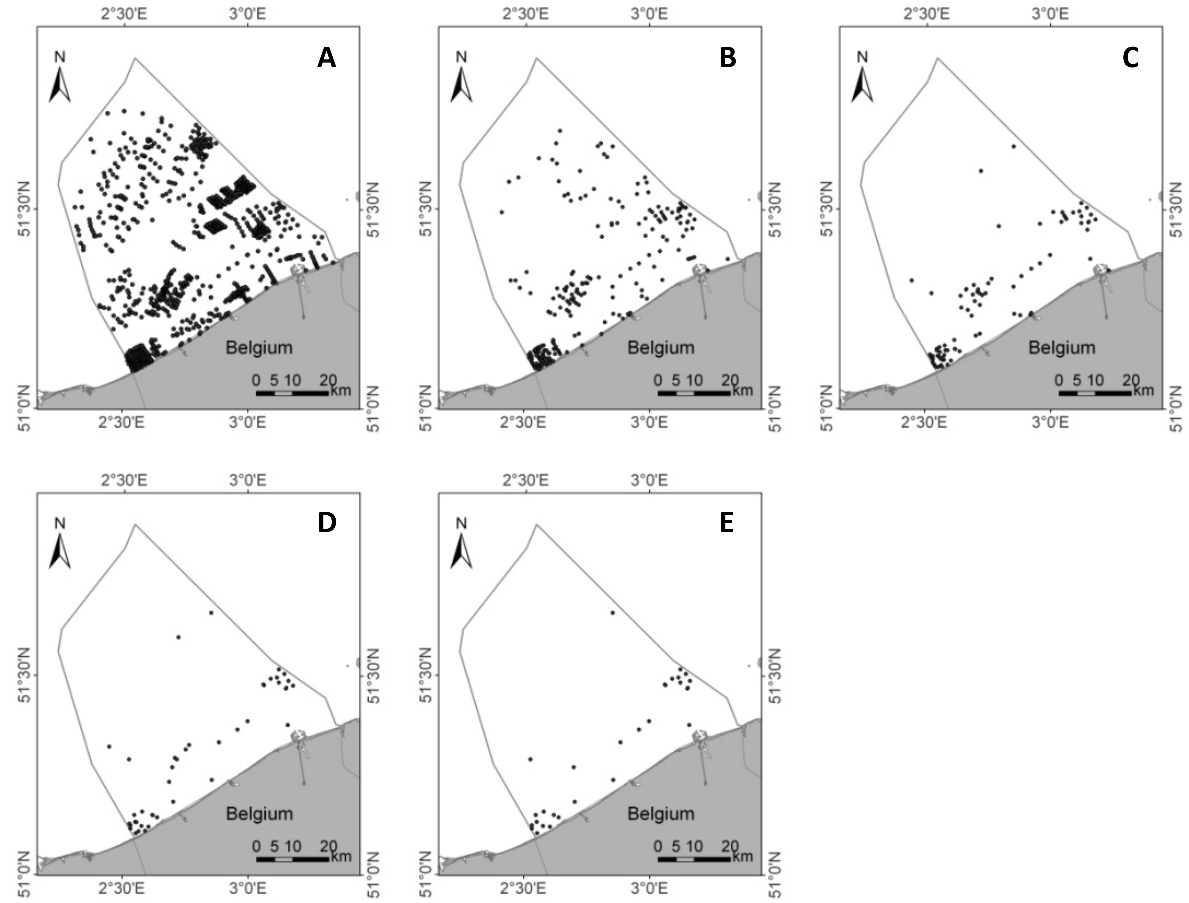

Fig. 7.1. Sampling stations (•) from the MacroDat database (A), stations where L. conchilega *has been observed (B),* L. conchilega *stations with at least 100 ind.m^{-2} (C),* L. conchilega *stations with at least 500 ind.m^{-2} (D) and* L. conchilega *stations with at least 1000 ind.m^{-2} (E).*

Environmental data

Fifteen environmental variables related to granulometry, the topography of the area and current properties were selected (Table 7.1). Chlorophyll *a* and total suspended matter were excluded from the analysis because these maps have no data values near the coastline, which would therefore result in maps without any prediction near the coast. Only those current properties concerning bottom currents and bottom shear stress were considered, since these variables may be of direct influence to *L. conchilega*.

	Variable	Abbreviation	Institute
Sediment related data	Median grain size	d50x	RCMG
	Gravel content	grav	RCMG
	Sand content (63 µm - 2 mm)	sand	RCMG
	Silt-clay content (0-63 µm)	mudx	RCMG
Topographical data	Water depth	dept	RCMG
	Slope of the sea bottom	slop	RCMG
	Bathymetric Position Index (1600 m range)	bpi2	RCMG
	Bathymetric Position Index (240 m range)	bpi3	RCMG
	Rugosity of the bottom	rugo	RCMG
	Orientation of the slope of the bottom	aspe	RCMG
Current properties	Minimum bottom shear stress	bsti	MUMM
	Mean bottom shear stress	bstm	MUMM
	Maximum bottom shear stress	bstx	MUMM
	Maximum current velocity at the bottom layer	umax	MUMM
	Average current velocity at the bottom layer	umea	MUMM

Table 7.1. Overview of the abiotic variables and their data source.

Modelling procedure

Maxent was used as a modelling algorithm (Addendum 1). The use of presence-only data can be justified, since *L. conchilega* aggregations have been considered to be ephemeral (Zühlke, 2001). Recent research showed that local individual aggregations can be short-lived, while large areas are persistently inhabited by *L. conchilega* over decades (Callaway *et al.*, 2010). Thus, absence does not necessarily mean that the habitat is not suitable for the species, but it may be the result of the potential ephemeral character of the species' distribution.

Preferential sampling cannot be a priori excluded as the sampling stations in the MacroDat database are not evenly spread across the region (Fig. 7.1). This may result in accepting models which are not significantly different from random and may be revealed by a randomisation test in which all the sampling stations are used at random to construct 'random species' models (Raes and ter Steege, 2007; Merckx *et al.*, 2011). The randomly selected coordinates are considered to be the locations where the 'random' species is found. In this way 999 random models were created and for each of these models, the area under the curve (AUC, Addendum 1) is calculated. A species model can be considered to be different from random when its corresponding AUC is significantly higher than the one-sided 95 % CI of the AUC-values of the random models. Since the number of stations influences the AUC-threshold value for a random model, this randomisation process was repeated for the four different threshold values.

When the randomisation test points out that the model is significantly different from random, the model is further fine-tuned by a backward and forward variable selection. This is done by a five-fold cross-validation, and the model with the highest average AUC is selected. The final model is then calculated by using all the data points and the restricted number of variables.

Results and discussion

Test for preferential sampling

The randomisations point out that the *Lanice* models are significantly different from random (Fig. 7.2). The difference between the AUC of the *Lanice* model is considerably higher for the random models of the total area, than for the random models selected from the actually sampled stations. This shows that there is actually a sample bias in the sampling stations: some areas were oversampled and others undersampled. Notwithstanding this sampling bias, the distribution of *Lanice conchilega* is significantly different from random for the four density thresholds. Thus, for each of the threshold densities a non-random habitat suitability model can be constructed. A forward and backward selection was performed. The forward selection for the L500 model selected only two variables related to water current properties: maximum current velocity at the bottom layer (umax) and average current velocity at the bottom layer (umea). These properties are also found by the backward modelling (see Fig. 7.3), but here the silt-clay fraction is also selected as an important variable, which is in accordance with previous research (Willems *et al.*, 2008). Therefore, only the backward selection is used.

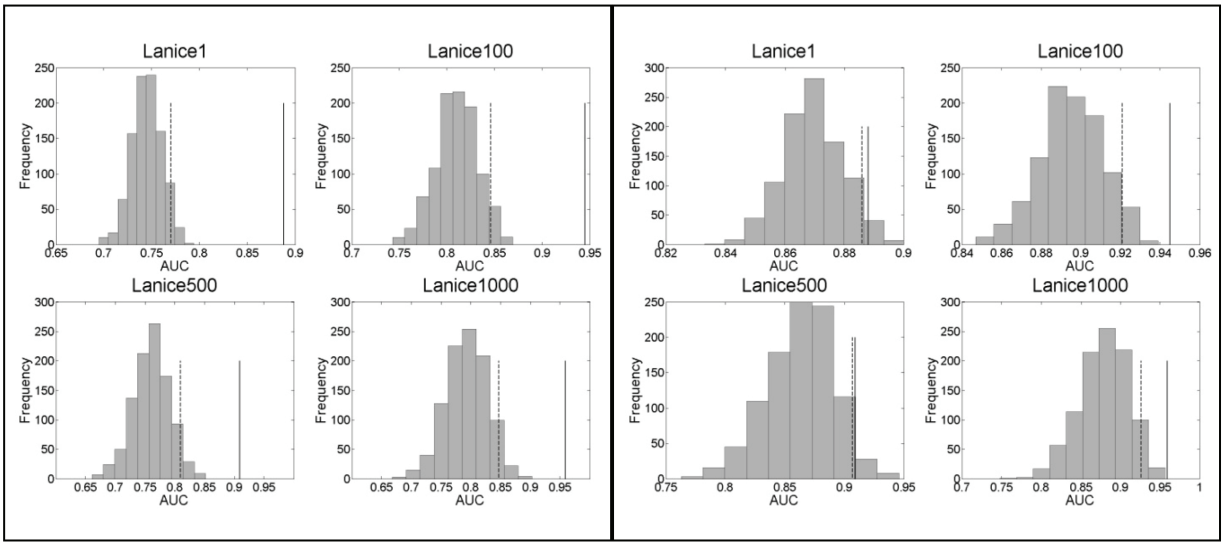

Fig. 7.2. Histograms of the randomisations of the four L. conchilega *densities: randomisations based on the total sampling area (left) and randomisations based on the actual sampling stations (right). The 95% quantile value (dotted line) and the AUC of the* Lanice *model (full line) are indicated.*

Contribution of the environmental variables

Maxent also supplies information on the relation between the species and its environment. *Lanice* is a cosmopolitan species and appears across a wide area of the Belgian Part of the North Sea (Van Hoey *et al.*, 2008) (Fig. 7.1). Since high densities of *Lanice* are of special interest, only the environmental factors contributing to the densities of at least 500 individuals.m^{-2} are considered (Table 7.2). Especially the silt-clay content and the maximum current velocity at the bottom layer are contributing to the model. Fig. 7.3 shows how the response of the *L. conchilega* model changes as each environmental variable is varied. Each of the curves represents a different model using only one environmental variable at the time. In this way the link between the selected variables and the logistic output of the model is demonstrated without interference of correlations between the environmental variables.

Variable	Abbreviation	% contribution
		Lanice500
Silt-clay content	mudx	70.8
Maximum current velocity at the bottom layer	umax	13.3
Bathymetric Position Index (240 m range)	bpi3	3.2
Slope	slop	7.8
Mean bottom shear stress	bstm	4.8

Table 7.2. Relative contributions of the environmental variables to the final Maxent model.

The response curves of the variables are shown in Fig. 7.3. Only the general pattern of the response curves is of importance, since the algorithm may still be slightly overfitting due to spatial autocorrelation. Presence-absence based habitat suitability modelling of *L. conchilega* highlighted the importance of the silt-clay fraction, the median grain size and the amount of coarse sediment fractions for the distribution of this species (Willems *et al.*, 2008). Field data pointed out that the highest *L. conchilega* densities are found in shallow fine sands (Van Hoey *et al.*, 2008) and shallow muddy sands (Van Hoey *et al.*, 2008). Depending on the classification, muddy sands contain between 10 and 50 percent silt-clay, or between 10 and 25% silt-clay (Long, 2006). The optimal silt-clay content according to our models ranges between approximately 0 and 20%, and falls thus in the range of muddy sands and fine sands. Our models also point out that the absence of silt-clay is not favoured either, because the model response drops to zero at zero silt-clay content (Fig. 7.3). At the highest silt-clay values the probability of occurrence slightly increases, which may be attributed to two stations in the area with high silt-clay content. Further research should indicate if the latter response is true or due to erroneous input data. Heuers *et al.* (1998) found that hydrodynamics are another important variable: the density of a *L. conchilega* assemblage increased significantly with increasing the flow velocity from 0.1 m.s^{-1} to 0.2 m.s^{-1}. Our results reveal the importance of the maximum current velocity at the bottom layer for the Lanice500 model. The model shows a positive relation between the probability to find *Lanice* aggregations and the maximum flow velocity, but only till values of about 0.6-

0.75 m.s^{-1} (optimum of umax in Fig. 7.3). When the maximum current velocity exceeds this value, the relationship turns into a negative one. Small slopes also show a positive relation with finding dense aggregates of the species and the ideal bathymetric position index (BPI) should be around zero. A BPI of zero indicates a flat area or an area with a constant slope. Since the slope should be small, zero BPI should be interpreted here as flat areas. The peak of the response curve of the slope near small values may be attributed to remaining overfitting, possibly due to spatial autocorrelation.

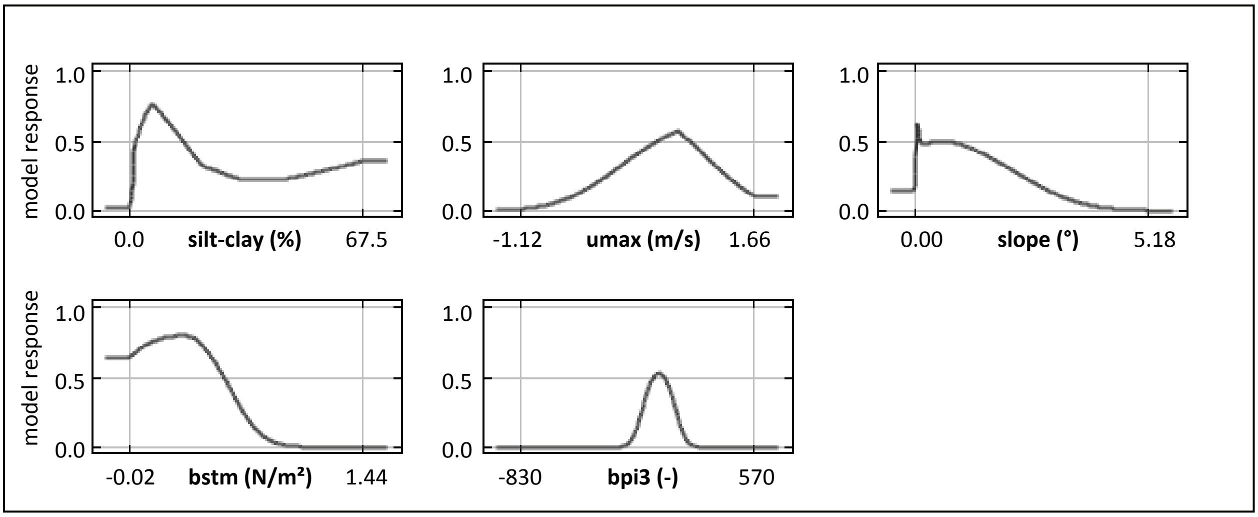

Fig. 7.3. Relation between the environmental variable and the logistic output of the Lanice conchilega models with densities of at least 500 individuals.m^{-2}. Each curve represents a model created using only the corresponding variable. See Table 7.1 for explanation of the abbreviations. The minimum and maximum values of the environmental variable are delineated by a vertical line. Before the minimum and after the maximum horizontal markers indicate the starting point and the endpoint of the curve. These markers do not have an ecological meaning.

Mapping of HSM of Lanice conchilega

The output map of the HSMs is continuous (Fig. 7.4). As expected, the areas with high probability of occurrence narrow down with increasing density threshold. By applying a threshold these maps can be converted to binary maps. Thresholds can be selected based on different claims: the sensitivity, the specificity and/or the purpose of the model. When it is important to map the area which encompasses the total distribution area of the species, a

Description	Logistic threshold	% of the total area predicted	% of *Lanice* in *Abra alba* area	% of *Abra alba* in *Lanice* area
10 percentile training presence	0.222	30	91%	51%
Maximum training sensitivity plus specificity	0.358	19	78%	68%

Table 7.3. Logistic thresholds and predicted area of Lanice conchilega *in comparison with* Abra alba *community.*

low threshold should be chosen. When high confidence of finding the species is required, a higher threshold can be applied. In this case, two thresholds have been applied (Table 7.3) 1) a threshold which results in a binary map where 90% of the presences are actually found in the predicted area (10 percentile training presence) and 2) a threshold resulting in a maximum value of sensitivity and specificity (see Addendum 1). Both methods are commonly used (Liu *et al.*, 2005; Weinsheimer *et al.*, 2010). Only for the latter threshold, which predicts a smaller fraction of the BPNS, a map is constructed (Fig. 7.4).

Comparison with the spatial distribution of the Abra alba *community*

The Belgian Part of the North Sea is a well-studied area and generally four macrobenthic communities are distinguished (Fig. 7.4); 1) the *Macoma baltica* community, characterised

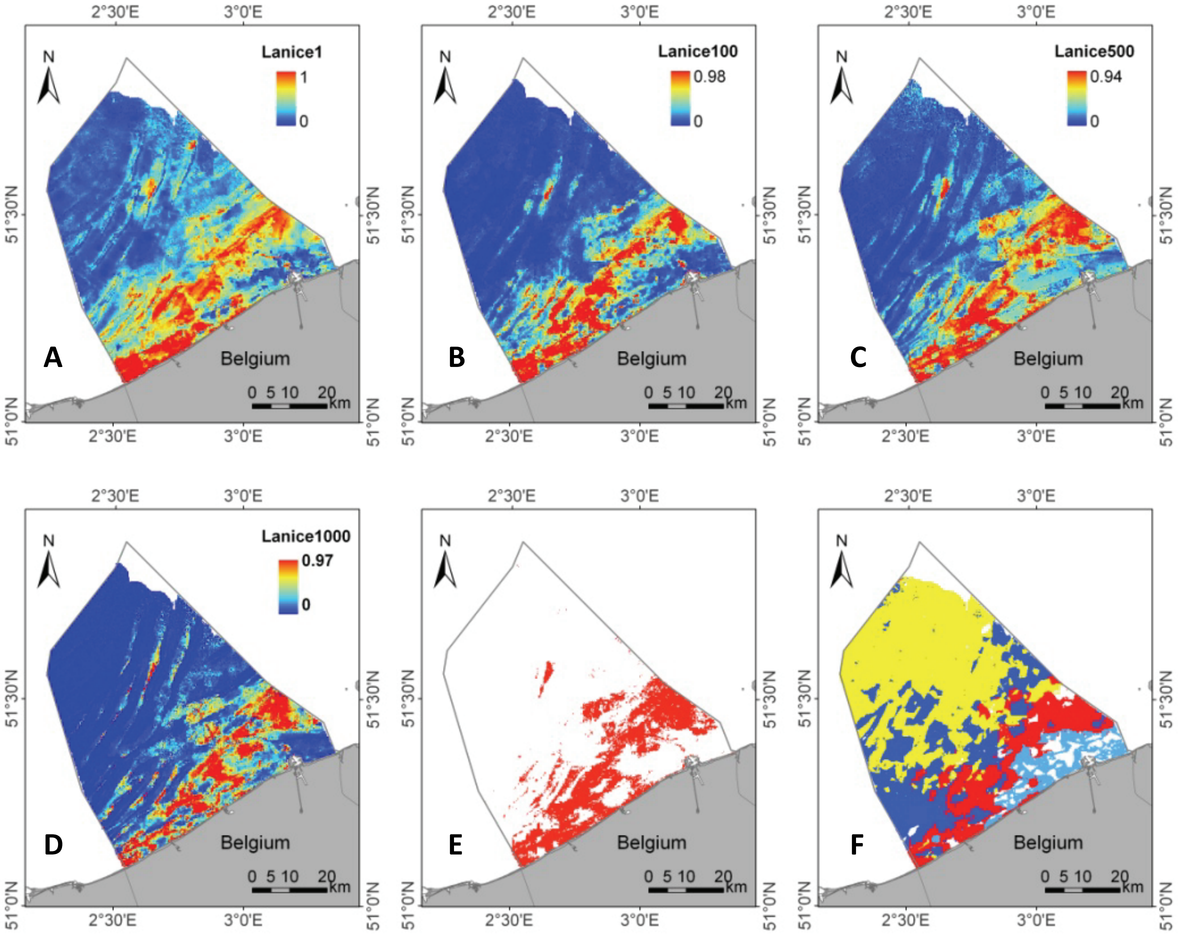

Fig. 7.4. Habitat suitability Models for different densities of Lanice conchilega*: presence (A), 100 ind.m*$^{-2}$ *(B), 500 ind.m*$^{-2}$ *(C), 1000 ind.m*$^{-2}$ *(D). Model of Lanice500 with 'maximum training sensitivity plus specificity' threshold (E) and delineation of the four main macrobenthic communities on the Belgian Part of the North Sea:* Macoma balthica *(pale blue),* Abra alba *(red),* Nephtys cirrosa *(blue) and* Ophelia limacina *(yellow) community (F) (from Degraer* et al.*, 2009).*

by quite high macrobenthic densities and low species richness. It is commonly found in muddy environments near the Eastern part of the Belgian coast; 2) the *Abra alba* community, characterised by both high macrobenthic densities and high species richness, and found in fine sand with high silt contents; 3) the *Nephtys cirrosa* community, typified by a generally low density and low diversity and found in environments with pure fine to median sand; 4) the *Ophelia limacina* (- *Glycera lapidum*) community holding very low densities and low diversity and which can be found in medium to coarse sands (Degraer *et al.*, 2003b; Van Hoey *et al.*, 2004).

The *Abra alba* community has previously been described as being of exceptional ecological importance because of its high macrobenthic abundance (6432 ind.m^{-2}) and diversity (30 species per sample of 0.12 m^2). This community tends to hold some unique species for the BPNS (Van Hoey *et al.*, 2004) and the bivalves present in the community may serve as food source to sea ducks (Degraer *et al.*, 1999). Moreover, *L. conchilega* has the capacity to double the biodiversity of the *A. alba* community (Van Hoey, 2006). Table 7.3 and Fig. 7.4 clearly indicate that aggregations of *Lanice* pattern overlap with the distribution of the *Abra alba* community (Fig. 7.4 and Table 7.3). For the first threshold, about 51% of the area where the Lanice500 model has a high probability is situated within the *Abra alba* region, but on the other hand the *Abra alba* community lies for about 91 % in the Lanice500 region (Table 7.3). When the threshold increases, a smaller area, but with a higher probability of Lanice500 is predicted. Still 78% of the Lanice500 model is found within the *Abra alba* community and 68% of the Lanice500 surface lies within the distribution area of the *Abra alba* community. The clear overlap between the two areas can thus only promote the relevance of *Lanice conchilega* for nature conservation.

ENSIS DIRECTUS (CONRAD 1843)

Introduction

Ensis directus (Atlantic jackknife, American jackknife clam or razor clam) is an edible bivalve, indigenous to the Atlantic coast of North-America (von Cosel, 1982). Probably, it has been introduced in Europe around 1978 as larvae in ballast water of a ship crossing the Atlantic (von Cosel, 1982). Since then it has spread across the Dutch and Belgian coast. The first observation in Belgium dates from 1987 (Kerckhof and Dumoulin, 1987). Nowadays, it is found along the entire Belgian coast where it forms dense banks (Kerckhof *et al.*, 2007). Since its expansive behaviour, questions arose about its potential distribution and harmful effect on the natural community. On the other hand, clam fisheries, prohibited in Belgium at the moment, have also shown interest in the distribution of this bivalve (Houziaux *et al.*, 2010).

The species tends to occur in high densities (i.e. bivalve banks) and densities of 1000-2000 ind.m^{-2} are not uncommon (Armonies and Reise, 1999; Tulp *et al.*, 2010). These *Ensis* banks have a patchy distribution, but patches are not permanent. The European *E. directus* populations show conspicuous events of mass mortality, mainly in late winter or early spring (Armonies and Reise, 1999). *Ensis directus* lives deep in the sediment and when in danger, it can retract fast into the sediment with its powerful foot down to a depth of 50 cm (Tulp *et al.*, 2010). It prefers muddy, fine sand with small amounts of silt (Beukema and Dekker, 1995; Kennish *et al.*, 2004) and is found in the intertidal and subtidal zones (Mühlenhardt-Siegel *et al.*, 1983; Swennen *et al.*, 1985). These sublittoral muddy fine sand sediments are also known to be the habitat of the diverse *Abra alba* community (Van Hoey *et al.*, 2004; Degraer *et al.*, 2008). Therefore, knowing the potential habitat of *E. directus* may indicate if a potential effect of *E. directus* on the *Abra alba* community is to be expected.

The objective of this study is (1) to identify the environmental factors related to the presence of *E. directus*; (2) to construct a habitat suitability model for the species and (3) to construct a map with the density distribution of the species.

Materials and methods

Research area

Ensis directus is found from the intertidal to water depths of about 20 to 30 m. Nowadays, native *Ensis* species are not longer found in water shallower than 20 m (Kerckhof F., pers. comm.). *Ensis* specimens found in coastal shallow waters can thus considered being *E. directus* (Tulp *et al.*, 2010). Therefore, only the area near the Belgian coast is a potential habitat for this species, and the research area is restricted to the 12-miles zone of the Belgian Part of the North Sea (Fig. 7.5).

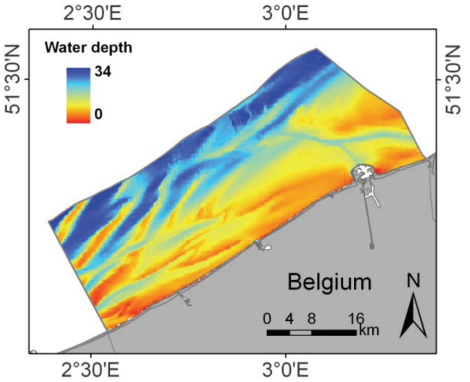

Fig. 7.5. Bathymetrical data of the 12-miles zone of the BPNS (Source: RCMG).

Ensis directus *data*

Two independent databases were analysed: 1) the MacroDat database (Degraer *et al.*, 2003a), completed with more recent data (2008), hereafter called MacroDat database and 2) data from a recent sampling campaign (2010) performed by the Management Unit of the North Sea Mathematical Models and the Scheldt estuary (MUMM), hereafter called the MUMM data.

The data from the MacroDat database was sampled with a Van Veen grab which has a penetration depth of about 7-10 cm (Degraer S., pers. comm.), and the MUMM data with a box corer which has a penetration depth of about 30 cm (Houziaux J.-S., pers. comm.). Since *E. directus* can easily withdraw up to 50 cm in the sediment (Tulp *et al.*, 2010), actual presence at a station cannot be ruled out when the species is not found in a sample.

The MacroDat database contains 869 sampling stations within the 12 miles zone (Fig. 7.6). In 201 stations *Ensis* specimens were found. The MUMM database holds data of 210 sampling stations, in 137 stations of these stations *E. directus* was found (Fig. 7.6). The database also contains information on the densities of two age classes, 1-year-old (D1) and older specimens (D2). The D1-class was found in 94 stations and the D2-class in 78 stations. Only in 37 stations both size classes were found. The sediment data at hand captures the surface sediment composition, where the younger specimens live. It should be noted that the survival of the older, deep-burrowing species, may be influenced by deeper sediment conditions for which no data are available. Therefore, at this stage the research focused only on the 1-year old specimens.

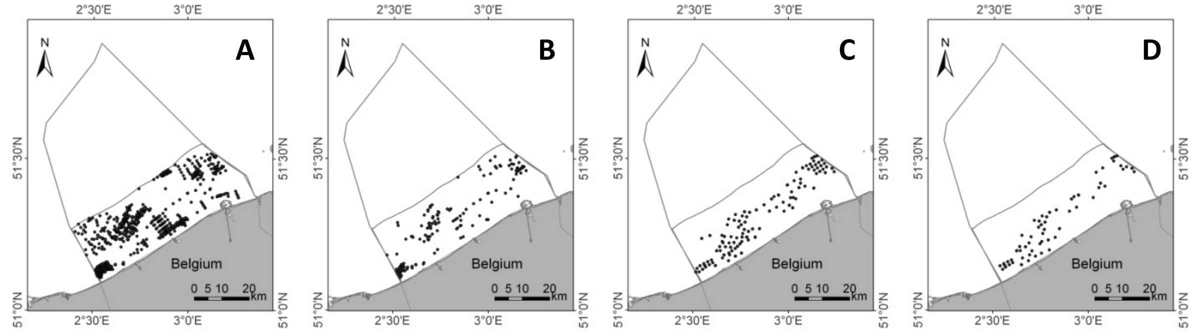

Fig. 7.6. Sampling stations (•) from the MacroDat database (A) within the 12-miles zone of the BPNS and stations where Ensis has been observed (B). Sampling stations from the MUMM database (C) and stations from the MUMM database where 1-year old specimens were found (D).

Environmental data

The environmental data has two sources: 1) sediment data collected simultaneously with the biological data, such as silt-clay, sand fraction and median grain size; 2) data extracted from exhaustive data maps (Table 7.4).

Modelling procedure

Habitat suitability modelling (HSM)

Ensis directus is an invasive species, which means that the species has not yet reached equilibrium in the area, and the absence of the species does thus not necessarily imply that the habitat is unsuitable for the species. Therefore, we used a presence-only modelling algorithm, namely Maxent.

Variable type	Variable	Abbreviation	Unit	Source
Biochemical	Average total suspended matter	tsme	$g.m^{-3}$	Belcolour
	Maximum total suspended matter	tsma	$g.m^{-3}$	Belcolour
	Minimum total suspended matter	tsmi	$g.m^{-3}$	Belcolour
	Average chlorophyll content	chme	$mg.m^{-3}$	Belcolour
	Maximum chlorophyll content	chma	$mg.m^{-3}$	Belcolour
	Minimum chlorophyll content	chmi	$mg.m^{-3}$	Belcolour
Hydrodynamical properties	Minimum bottom shear stress	bsti	$N.m^{-2}$	MUMM
	Mean bottom shear stress	bstm	$N.m^{-2}$	MUMM
	Maximum bottom shear stress	bstx	$N.m^{-2}$	MUMM
	Size of the residual currents	mcur	$m.s^{-1}$	MUMM
	Maximum depth-averaged current velocity	mmax	$m.s^{-1}$	MUMM
	Magnitude of the residual transports	mtra	$m.s^{-1}$	MUMM
	Residual currents	rcur	$m.s^{-1}$	MUMM
	Residual transports	rtra	$m.s^{-1}$	MUMM
	Tidal amplitude	tamp	m	MUMM
	Maximum current velocity at the bottom	umax	$m.s^{-1}$	MUMM
	Average current velocity at the bottom	umea	$m.s^{-1}$	MUMM
Topographic properties	Water depth	dept	m	RCMG
	Slope of the sea bottom	slop	°	RCMG
	Bathymetric Position Index (1600 m range)	bp20	-	RCMG
	Bathymetric Position Index (240 m range)	bp13	-	RCMG
	Rugosity of the bottom	rugo	$m^2.m^{-2}$	RCMG
	Orientation of the slope of the bottom	aspe	°	RCMG
Sediment	Median grain size	d50x	µm	RCMG
	Gravel content	grav	weight%	RCMG
	Sand content (63 µm - 2 mm)	sand	%	RCMG
	Silt-clay content (0-63 µm)	mudx	%	RCMG

Table 7.4. Environmental variables and their data source. See p.v for more information about the source of the variables.

The MacroDat dataset contains many data points, which has a strong effect on the calculation time of the modelling algorithm. Therefore, a preliminary variable selection was carried out based on the Spearman rank correlation between the environmental variables and a jackknife test in Maxent. If the Spearman rank correlation between two variables was larger than 0.8, the variable performing the worst in the jackknife test was removed. The jackknife test was carried out in Maxent to identify those environmental variables with the lowest gain when used in isolation. The MUMM database holds less samples and a preliminary variable selection was not performed.

The presence of preferential sampling was checked by applying a randomisation test: 499 random models were created with locations sampled from the actually sampled stations. When preferential sampling is not evident from the data, the model is further refined by applying a feature and variable selection. Maxent applies default different features: linear, quadratic, product, threshold and hinge features. Complex features may enhance overfitting, therefore a feature selection may result in a more realistic relation between the environmental variables and the output of the HSM. The variable and feature selection was done by a five-fold cross-validation.

Spatial autocorrelation may enlarge the AUC of the test dataset and may lead to overfitting and less realistic models. To evaluate the influence of spatial autocorrelation, we applied this model optimisation procedure for three distances between the training dataset and the test dataset: 0 km, 1 km and 5 km. The resulting models will be referred to as Ensis0km, Ensis1km and Ensis5km.

Density map

The HSM offers an indication of where the species potentially can be found, without differentiation between presence in high or low densities. However, regions with high densities of the species may have the strongest influence on the indigenous community, or may function as feeding areas of sea birds or for fisheries. Therefore, a density map was constructed as well. To construct such a map, geostatistics were applied. This involved two kriging algorithms: (1) ordinary kriging (OK) and (2) regression kriging (RK) (see Addendum 1) combined with a generalised linear model (GLM). The performance of both modelling techniques is compared by an independent test set, containing 20% of the data. This test set is solely used at the completion of the analysis. Different quality parameters were calculated to estimate the quality of the model output, compared to the real values in the test set: the mean estimation error (MEE), the root mean squared error (RMSE), the Pearson product-moment correlation coefficient (Pearson), the Spearman rank correlation coefficient (Spearman) and the mean absolute estimation error (MAEE) (see Chapter 4 for equations of these quality parameters).

Results and discussion

Habitat suitability modelling

Test for preferential sampling

The randomisation exercise (Fig. 7.7) points out that the sampling strategy of the MacroDat database is less biased than the MUMM data. This makes sense, since the recent MUMM sampling campaign was targeted towards *Ensis directus*. The sampling stations were selected in such a way that there was a high probability to find *E. directus*. Areas where the species is less likely to be found are thus undersampled. The AUC of the *Ensis* model is 0.93, which is generally considered to be an excellent model (Parolo *et al.*, 2008), but notwithstanding this high AUC the HSM cannot be distinguished from a random model due to preferential sampling. Therefore, we will only focus on the HSM model derived from the data from the MacroDat database.

The three models with different distances between training and test data, i.e. Ensis0km, Ensis1km and Ensis5km, were further refined with a variable and feature selection. The final Ensis0km model uses all features and 10 environmental variables in the final model. This results in overly complex relations between the variable and the model output. For instance two or three optimum values for the minimum bottom shear stress and median grain size are found. The reason why such an overly complex model is selected by cross-validation can be addressed to overfitting and spatial autocorrelation. Since there are no restrictions to the distance between the samples in the training and the test set, it is likely that the samples in the test and training set are spatially close to each other. Therefore the values of both the environmental variables and the output may be very similar for both datasets due to spatial autocorrelation. This means that, notwithstanding the cross-validation test, the model can still be overfitted because of the similarity between the test and the training set. Hence, the Ensis1km and Ensis5km model will suffer less from overfitting. Indeed, the 1 and 5 km model are much more realistic, as respectively 5 and 3 variables and only linear and quadratic features are selected (Table 7.5, Fig. 7.8).

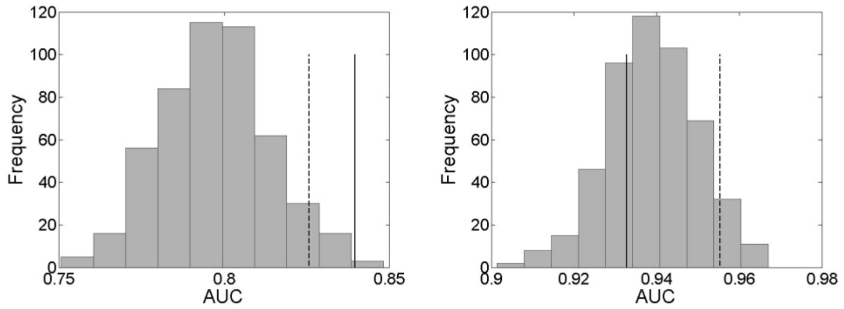

Fig. 7.7. Histograms of the randomisations of the MacroDat (left) and the MUMM database (right). The randomisations are based on the actual sampling stations. The 95% quantile value (dotted line) and the AUC of the Ensis *model (full line) are indicated.*

Contribution of the environmental variables

Both models, Ensis1km and Ensis5km, select bottom shear stress, water depth and sand fraction as the most important factors. For the Ensis1km model the minimum and maximum chlorophyll content are selected as well. Previous research pointed out that the American Jack knife clam is an opportunistic species, with little requirements regarding its environment. It prefers wave- and current-swept clean sands (Beukema and Dekker, 1995) with small amounts of silt (Kennish *et al.*, 2004), but it can also be found in muddy or coarse sediments (Armonies and Reise, 1999) and can thus be independent of sediment characteristics (Dauvin *et al.*, 2007). *E. directus* however has a limited tolerance to hypoxia and will thus avoid reduced sediments (Schiedek and Zebe, 1987). The positive relation between the sand fraction and *E. directus* is found in both HSM models (Fig. 7.8). Silt-clay was not selected as a variable. The preference for moving sands (Kenchington *et al.*, 1998) and strong currents are not advocated by the model; the model indicates that the maximum bottom shear stress should be limited to about 4 N.m^{-2} while a shear stress above 5 N.m^{-2} corresponds to the threshold of sand transport (Mangelsdorf *et al.*, 1990). Likely areas for colonisation are subtidal and intertidal areas (ICES, 2005; Ovcharenko and Gollasch, 2009). The models do not favour near shore areas, which is probably because there is no environmental data available in the intertidal area. Intermediate water depths between 12 and 23 m are optimal according to the models. This is partially in agreement with observations by Armonies and Reise (1999), who found that sublittoral populations prefer water depths of 18 m and more. The influence of chlorophyll *a* on the Ensis1km seems contradictory: the species is favoured by low minimum chlorophyll *a* values and high maximum chlorophyll *a* values. This could mean that the species is preferably found in areas with annually strongly fluctuating chlorophyll *a* values, however, there are no literature sources supporting this.

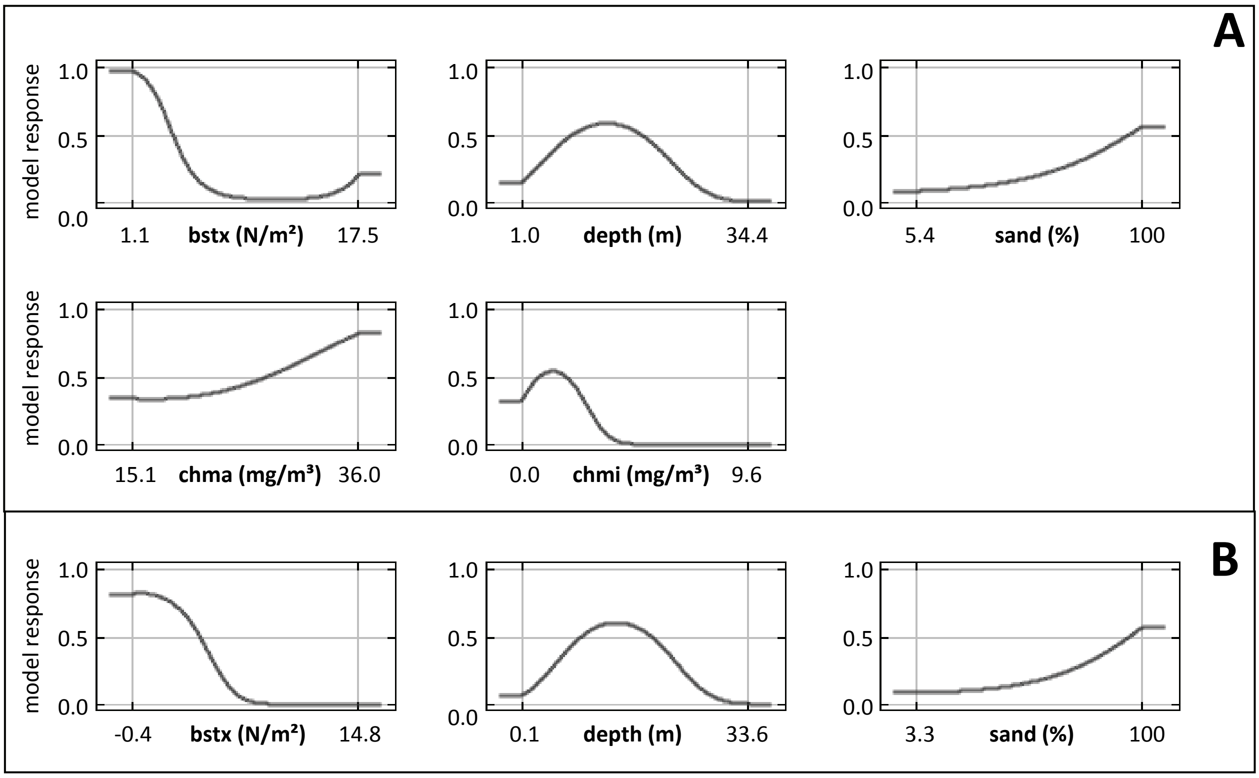

Fig. 7.8. Relation between the environmental variables and logistic output of the Ensis1km (A) and Ensis5km model (B). Each of the curves represents a model created using only the corresponding variable. See Table 7.4 for explanation of the abbreviations. The minimum and maximum values of the environmental variable are delineated by a vertical line. Before the minimum and after the maximum horizontal markers indicate the starting point and the endpoint of the curve. These markers do not have an ecological meaning.

The selected variables do not concord very well with findings in literature. This can be explained by two reasons: First, a lot of the data in literature is collected in intertidal areas. In the present study, information on the intertidal is lacking. Secondly, all observations were taken into account, but *E. directus* is an opportunistic species which may occur in many different habitats. Its presence is therefore only slightly regulated by the environmental variables. However, it is possible that high densities of *E. directus* can only thrive in specific conditions. Therefore, by analogy to Chapter 6 and the HSM developed for *Lanice conchilega*, creating HSM for the species based on density thresholds could reveal the environmental variables which relate to high densities of the species.

Variable	Abbreviation	% contribution	
		Ensis1km	Ensis5km
Maximum bottom shear stress	bstx	27.4	39.7
Water depth	dept	22.8	35.9
Sand Fraction	sand	15.9	24.4
Maximum chlorophyll content	chma	29.9	
Minimum chlorophyll content	chmi	4.1	

Table 7.5. Relative contributions of the environmental variables to the final Maxent model.

Comparison with the spatial distribution of the Abra alba *community*

The resulting map of the Ensis1km model is shown in Fig. 7.9. This map can be transformed into a binary map by choosing a logistic threshold. Here, the '10 percentile training presence' threshold was used. This is a widely used threshold (Weinsheimer *et al.*, 2010). This results in a map which covers about 45% of the investigated area. It reveals the same pattern as the *Abra alba* community (Fig. 7.4) and interactions for space and food might be expected. However, strong interactions with the indigenous fauna have not yet been established: along the Island of Sylt (North Sea) (Armonies and Reise, 1999) only one negative relation with *Cerastoderma edule*, the common cockle, was observed, while the other infaunal species showed no convincing interaction with *E. directus*. Other literature sources expect little or no interaction since *E. directus* appears in high densities in lower intertidal sand flats and offshore sand banks which have general low macrobenthos densities and these poorly populated areas may thus represent an unoccupied niche for an opportunistic species such as *E. directus* (Beukema and Dekker, 1995; von Cosel, 2009).

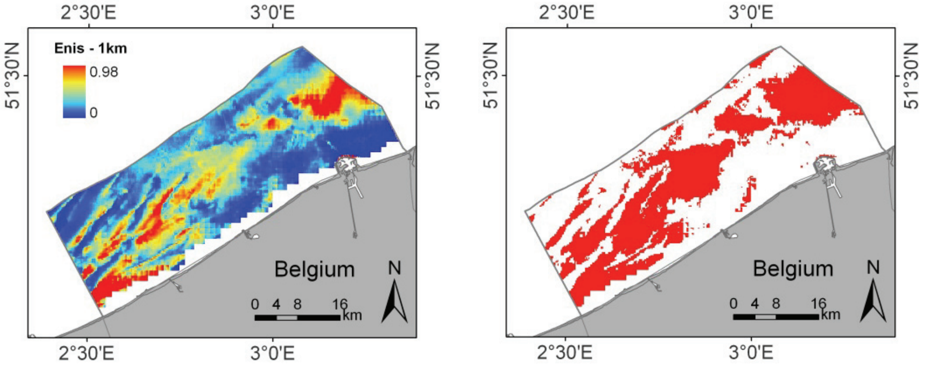

Fig. 7.9. Resulting maps of the Ensis1km habitat suitability model and the thresholded model ('10 percentile training presence' threshold) .

Density map

The habitat suitability map (Fig. 7.9) indicates whether the habitat is suitable for the species but it does not reveal information about expected densities. Therefore, a density map for *E. directus* was also constructed. Since there were large differences between the densities, and only a few stations show high densities, the data was log-transformed. Two techniques were applied: OK and RK. For RK, a linear model was constructed but the relation between the log-density of the young cohort (D1) and the environmental variables was not strong. Two variograms were constructed: one for the OK and one for RK, based on the residuals of the linear model. The most important parameters of the variograms can be found in Table 7.6. The range of the OK variogram is 5.4 km, thus within this range there is a spatial dependency between the density of the samples. The sill and the nugget of the variogram of RK are larger than of OK, which is unusual. In fact, the linear model should explain a portion of the variation in the data, which would result in a decreasing sill. This observation is also supported by the results in Table 7.6. Most of the quality parameters, except for the Spearman rank correlation coefficient, perform better for OK. Thus, OK performs better than RK. Hence, the vicinity of high densities can explain better the presence of high densities than the environmental variables, and no strong relationship is found between the environmental factors and the density of the young cohort of *Ensis*. Spatial autocorrelation and spatial interpolation explains the densities of *E. directus* better than the environmental variables.

	Variogram parameters			Model validation parameters				
	Nugget	sill	Range (km)	MEE	RMSE	Pearson	Spearman	MAEE
Ordinary kriging	0.24	1.50	5.4	-0.02	0.59	0.83	0.84	0.45
Regression kriging	0.90	1.80	9.4	0.03	0.70	0.76	0.87	0.61

Table 7.6. Variogram parameters and model validation parameters calculated between the predicted and the real values of the test set.

With the data at hand it is not possible to make an area covering map of the *Ensis* densities (Fig. 7.10). The highest densities were found off coast and these densities show a patchy distribution. The stations appear as spots on the variance map, since it is assumed that the variance is smallest at the sampling location. To construct relevant maps which give an indication of the species density, we suggest constructing habitat suitability models based on density thresholds, as was done for *L. conchilega* and some nematode species (Chapter 6). Therefore, size and density thresholds should be identified for which an impact on the surrounding benthos is expected or thresholds which are relevant to fisheries or sea birds.

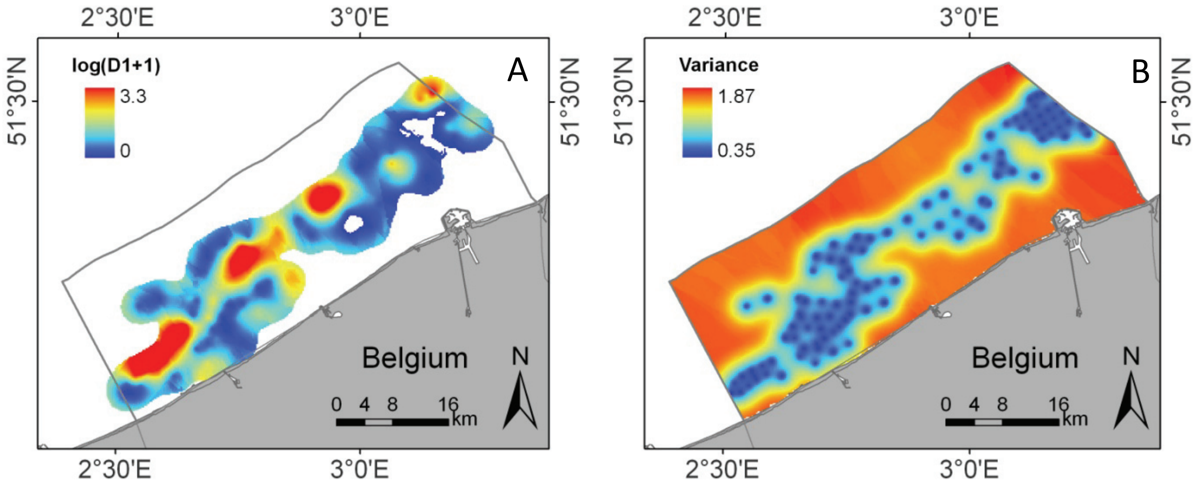

Fig. 7.10. Map of log(D1+1), with D1 the density of the 1-year old cohort (ind.m^{-2}) (A). This map is restricted to a variance smaller than 1.5. On the right side the variance map is shown (B).

COMPARISON BETWEEN *LANICE CONCHILEGA*, *ENSIS DIRECTUS* AND *ABRA ALBA* COMMUNITY

To get an overall idea of the area shared by *Lanice conchilega*, *Ensis directus* and the *Abra alba* community, all the maps were represented as binary maps for the 12-miles zone (Fig. 7.11). For the Lanice500 model the 'Maximum training sensitivity plus specificity' threshold was chosen, while for the *Ensis* model the '10 percentile training presence' threshold was chosen. These two thresholds result in comparable fractions of the totally investigated area, namely 40% for *L. conchilega*, 45% for *E. directus* and 38% for the *Abra alba* community. These individual areas overlap considerably; the Lanice500 model and the *Abra alba* community have the largest area in common, namely three fourth of their area or 29% of the total area. The overlap between *Ensis* and both other species takes up about two third of *Ensis*' area. The three models overlap in 21% of the total area or for about 50% of their individual space. Thus, it is reasonable to believe that *Ensis* may affect the indigenous community since they potentially share a considerable amount of space.

The potential effect of *Ensis directus* on the *Abra alba* community and on *Lanice conchilega* is poorly studied. However, existing data does not report effects: the introduction of *E. directus* does not affect the settlement of high densities of *L. conchilega* (Ghertsos *et al.*, 2000) and *E. directus* has been reported to be short-lived at certain sites within the *Abra alba* community (Ghertsos *et al.*, 2000). However, the long-term influence of *E. directus* on the indigenous communities, is largely unknown and should be carefully monitored since the *Abra alba* community is the most species rich soft bottom macrobenthic community on the Belgian Part of the North Sea.

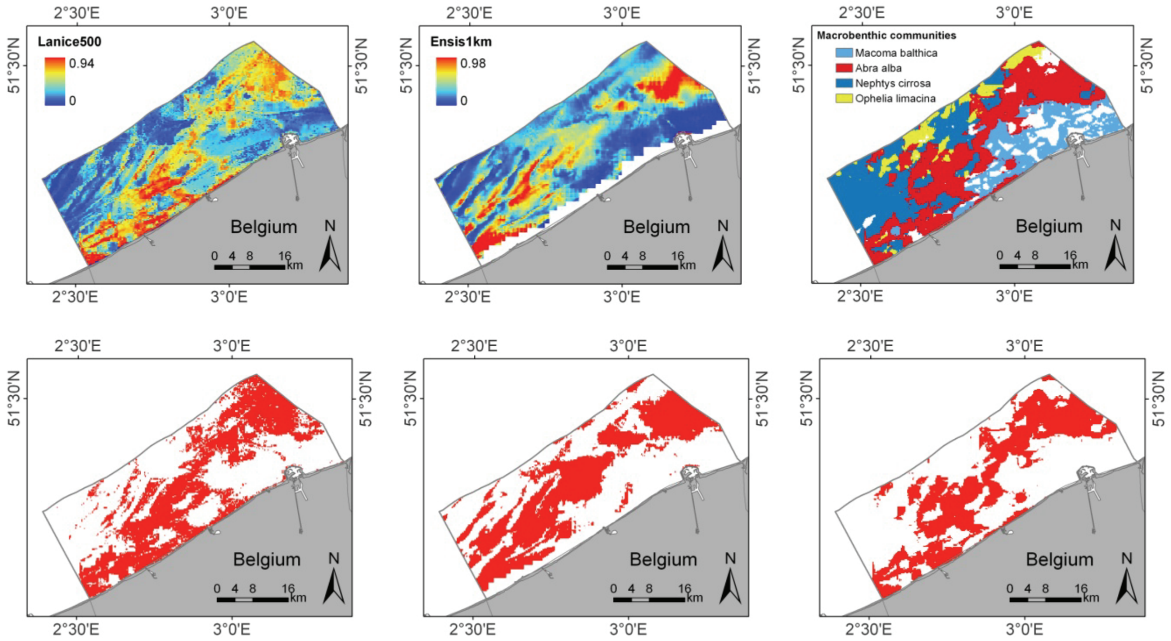

Fig. 7.11. Habitat suitability models of Lanice conchilega *(500 ind.m^{-2}),* Ensis directus *(1 km) and the four macrobenthic communities. Below, the reduction of these maps to binary maps representing* Lanice conchilega *(500 ind.m^{-2}, logistic threshold 0.358),* Ensis directus *(1 km, logistic threshold 0.206) and the* Abra alba *community.*

CONCLUSIONS

The methods developed for nematode species, are readily applicable to macrobenthic species. Potential pitfalls such as overfitting, spatial autocorrelation and preferential sampling should be countered whenever spatial data is analysed. Cross-validation and splitting datasets in subsets which are spatially separated help in addressing these problems and in selecting parsimonious models. Depending on the purpose of the model, the modelling technique can be adapted, i.e. modelling density thresholds instead of presences whenever high densities are important or when modelling an opportunistic species. Comparing the modelling results with previous research remains essential for three reasons: 1) if previous research confirms the results of the model, this strengthens the results of both the previous findings and the models; 2) this information may reveal potential problems with the modelling technique, for instance unpredicted regions because of missing data, unrealistic relations between environmental variables and the species because of overfitting or preferential sampling; 3) new insights to science and potential research topics may be revealed.

ACKNOWLEDGMENTS

This research is funded by the Fund for Scientific Research(FWO) of the Flemish Government (FWO07/ASP/174). The authors wish to thank all the data providers! The environmental data was gathered from different institutes: ESA and MUMM/RBINS are acknowledged for providing and processing MERIS data (chlorophyll *a* and TSM data, http://www.mumm.ac.be/BELCOLOUR), the Renard Centre of Marine Geology (RCMG, http://www.rcmg.ugent.be) of Ghent University and the Hydrographic Service of the Royal Netherlands Navy and the Directorate-General of Public Works and Water Management of the Dutch Ministry of Transport, Public Works and Water Management for the oceanographic and sedimentological data. The *Lanice conchilega* modelling was conducted within the Westbanks Project which is supported by the Belgian Science Policy (BELSPO): SSD Science for sustainable Development General Coordination and for the project 'Studie betreffende het opstellen van een lijst van potentiële Habitatrichtlijngebieden in het Belgische deel van de Noordzee' under the authority of Federale Overheidsdienst Volksgezondheid, Veiligheid van de Voedselketen en Leefmilieu, Directoraat-generaal Leefmilieu. The *Ensis directu*s modelling was performed as part of the Ensis project, financed by the Federal Science Policy.