CHAPTER 5

NULL MODELS REVEAL PREFERENTIAL SAMPLING, SPATIAL AUTOCORRELATION AND OVERFITTING IN HABITAT SUITABILITY MODELLING

Adapted from: Merckx, B., Steyaert, M., Vanreusel, A., Vincx, M., Vanaverbeke, J., 2011. Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. Ecological Modelling 222, 588-597.

CHAPTER 5

NULL MODELS REVEAL PREFERENTIAL SAMPLING, SPATIAL AUTOCORRELATION AND OVERFITTING IN HABITAT SUITABILITY MODELLING

ABSTRACT

Nowadays, species are driven to extinction at a high rate. To reduce this rate it is important to delineate suitable habitats for these species in such a way that these areas can be suggested as conservation areas. The use of habitat suitability models (HSMs) can be of great importance for the delineation of such areas. In this study Maxent, a presence-only modelling technique, is used to develop HSMs for 223 nematode species of the Southern Bight of the North Sea. However, it is essential that these models are beyond discussion and they should be checked for potential errors. In this study we focused on two categories (1) errors which can be attributed to the database such as preferential sampling and spatial autocorrelation and (2) errors induced by the modelling technique such as overfitting. In order to quantify these adverse effects thousands of nulls models were created. The effect of preferential sampling (i.e. some areas where visited more frequently than others) was investigated by comparing null models sampling the actual sampling stations with null models sampling the entire mapping area (Raes and ter Steege, 2007). Overfitting is exposed by a fivefold cross-validation and the influence of spatial autocorrelation is assessed by separating test and training sets in space. Our results clearly show that all these effects are present: preferential sampling has a strong effect on the selection of non-random species models. Cross-validation seems to have less influence on the model selection and spatial autocorrelation is also strongly present. It is clear from this study that predefined thresholds are not readily applicable to all datasets and additional tests are needed in model selection.

Keywords:

Maxent, null models, preferential sampling, spatial autocorrelation, overfitting, Nematoda

INTRODUCTION

Biodiversity and the conservation of species is a major concern in ecology nowadays. Species are driven to extinction at a high rate due to overexploitation, climate change and resource consumption (Butchart *et al.*, 2010). Not only the terrestrial space is fragmented and confronted with disappearing natural habitats, also the natural habitats in the oceans are endangered (Hoegh-Guldberg and Bruno, 2010).

The sea bottom is under peril due to bottom trawling, aggregate extraction, dredging and dumping. These habitat disturbances may threaten species to disappear. For conservation strategies, it is important to investigate the habitat preferences of species, and particularly of rare species to delineate and protect suitable habitats for these species.

Habitat suitability models (HSMs) can be a tool in protecting and conserving species (Rodriguez *et al.*, 2007). However, it is of major importance that these models are beyond discussion. These models need to be tested profoundly before they can be considered for conservation purposes. Several potential pitfalls need to be circumvented during modelling: spatial autocorrelation, preferential sampling, overfitting due to the use of oversized models and the use of redundant information (Pearson *et al.*, 2007; Parolo *et al.*, 2008). Different validating techniques can be applied during the modelling process: cross-validation is known to cope with overfitting, while null models help in identifying models significantly different from random. The latter approach also helps in identifying preferential sampling in datasets (Raes and ter Steege, 2007). The influence of spatial autocorrelation on the performance of the models can be tested by subdividing the data in spatially separated subsets which are in our case at least 5 or 10 km apart. In this study, we combine cross-validation and the null model approach to identify those models which are truly significantly different from random and not subject to preferential sampling, overfitting and spatial autocorrelation.

These modelling techniques are applied to a dataset of free-living marine benthic nematodes from the Southern Bight of the North Sea. Nematodes are usually the dominant taxon within the meiofauna, comprising metazoans passing through a 1 mm mesh sieve but retained on a 38 µm mesh sieve. These free-living roundworms represent the highest metazoan diversity in many benthic environments in terms of species numbers (Heip *et al.*, 1985). Owing to their interstitial life style, properties of the sediment, such as grain size distribution, the silt-clay fraction and food availability have a strong influence on the composition of nematode assemblages (Heip *et al.*, 1985; Vanreusel, 1990; Vincx, 1990; Merckx *et al.*, 2009, 2010). Nematode communities seem to be resilient to disturbance and their restoration occurs easily after temporal, low impacts (Kennedy and Jacoby, 1999; Schratzberger *et al.*, 2002), making them a perfect community to model based on long term environmental and full coverage data.

MATERIALS AND METHODS

Data

The research area, with a total surface of about 18 000 km^2, is situated in the Southern Bight of the North Sea, near the Belgian and the Dutch coastal area (latitude: 51°6'2'' to 52°59'19'' N; longitude: 2°14'39'' to 4°30'43'' E) (Fig. 5.1). The seafloor is not at all homogeneous in this area; it is characterised by sand dunes and a wide range of sediment types, varying from muddy to sandy environments (Lanckneus *et al.*, 2002). The coastal zone is characterised by a high amount of total suspended matter, chlorophyll *a* and silt-clay fraction, especially near the Belgian coast.

The nematode data were retrieved from the MANUELA database. Within the EU Network of Excellence MarBEF, MANUELA is a Research Project focusing on the meiobenthic assemblages. The MANUELA database was compiled capturing the available data on meiobenthos on a broad European scale (Vandepitte *et al.*, 2009). For this paper the area of research was restricted to the Southern Bight of the North Sea since full coverage environmental maps were available for this region.

The environmental variables were retrieved from maps acquired by remote sensing and maps interpolated from data sampled in the field.

The first group of maps summarises data on total suspended matter and chlorophyll *a* in the water column (Park *et al.*, 2006). The data is collected by remote sensing by the MERIS spectrometer on board of the Envisat satellite of the ESA. Eighty chlorophyll *a* maps and 90 total suspended matter maps were gathered during the time frame 2003-2005. These maps were reduced to three biologically relevant maps revealing the minimum, maximum and average values. This data reduction technique is often applied in ecological modelling (Loiselle *et al.*, 2008; Cunningham *et al.*, 2009; Echarri *et al.*, 2009). Satellite data are restricted to the water column but are of relevance for seafloor inhabiting organisms as sedimentation and degradation of chlorophyll *a* and total suspended matter enrich the bottom organic matter (Druon *et al.*, 2004). This input of organic matter is known to influence nematodes directly as it serves as a food source (Vanaverbeke *et al.*, 2004b; Franco *et al.*, 2008) or indirectly as microbial degradation often results in oxygen stressed sediments (Graf, 1992) which can have a strong adverse effect on nematodes (Steyaert *et al.*, 1999).

The second group contains maps derived from point sampling at sea. It comprises data on sediment characteristics, such as median grain size and the silt-clay fraction, and bathymetry. These maps were supplied by the Renard Centre of Marine Geology, Ghent University (Verfaillie *et al.*, 2006) and TNO Built Environment and Geosciences-Geological Survey of the Netherlands. The bathymetrical data were provided by the Ministry of the Flemish Community Department of Environment and Infrastructure, Waterways and Marine Affairs Administration and completed with data from the Hydrographic Service of the Royal Netherlands Navy and by the Directorate-General of Public Works and Water Management

of the Dutch Ministry of Transport, Public Works and Water Management. The silt-clay fraction and the median grain size are important factors determining the meiobenthic community (Heip *et al.*, 1985; Steyaert *et al.*, 1999; Vanaverbeke *et al.*, 2002; Merckx *et al.*, 2009). Depth in shallow waters does not directly affect the nematode community, but it modifies effects of other factors such as trophic conditions, sediment properties and current properties. An overview of the range of the environmental data in the dataset is shown in Table 5.1.

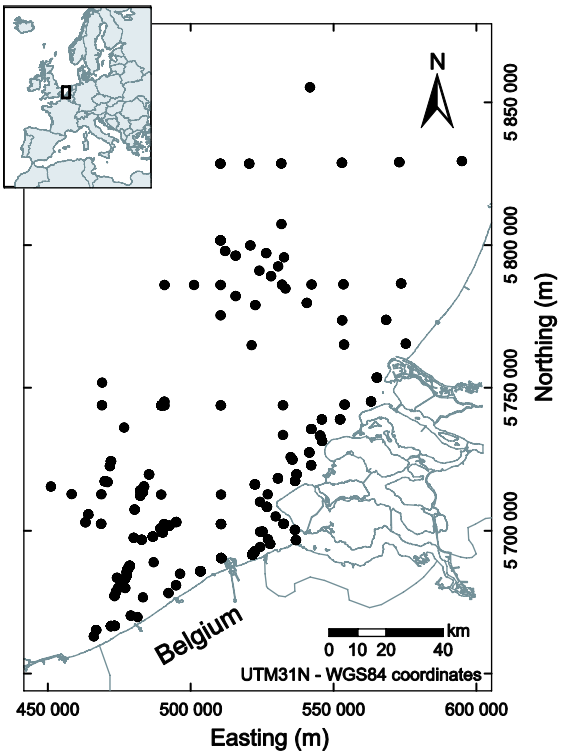

Fig. 5.1. Study area and location of the sampling stations (•).

Variable	Unit	Minimum	Maximum	Median
Silt-clay content	%	0	84	0.053
Total suspended matter (average)	$g.m^{-3}$	1	24	2.6
Total suspended matter (maximum)	$g.m^{-3}$	2.3	66	7.3
Total suspended matter (minimum)	$g.m^{-3}$	0.2	14	0.8
Chlorophyll *a* (average)	$mg.m^{-3}$	1.3	26	3.2
Chlorophyll *a* (maximum)	$mg.m^{-3}$	2.7	39	12
Chlorophyll *a* (minimum)	$mg.m^{-3}$	0.04	20	1.1
Depth of the water column	m	-1.3	53	26

Table 5.1. Range and median values of the environmental variables of the maps.

Habitat suitability modelling

Numerous modelling techniques and algorithms exist to investigate relationships between species and their environment in order to map their spatial distribution (Guisan and Zimmermann, 2000; Guisan and Thuiller, 2005). In several independent cases, the use of Maxent resulted in good predictive models compared to other presence-only models (Elith *et al.*, 2006; Hernandez *et al.*, 2006, 2008; Hijmans and Graham, 2006; Pearson *et al.*, 2007; Sergio *et al.*, 2007; Carnaval and Moritz, 2008; Ortega-Huerta and Peterson, 2008; Wisz *et al.*, 2008; Benito *et al.*, 2009; Roura-Pascual *et al.*, 2009). The reliability of the results of Maxent has been confirmed by its good capacity to predict novel presence localities for poorly known species (Pearson *et al.*, 2007). Besides the good predictive qualities of the technique, it has several other advantages: (1) it requires only presence data. For nematode data, this is an advantage as species absence is never certain since only a subsample of these inconspicuous organisms is identified in ecological research. (2) Overfitting can be avoided by using a regularisation mechanism (Phillips *et al.*, 2006). (3) Maxent is a generative approach, rather than discriminative, which can be an inherent advantage when the amount of training data is limited (Phillips *et al.*, 2006). This allows using the technique with as little as 5 sampling points (Pearson *et al.*, 2007). (5) It is possible to computerise the calculation of thousands of HSMs by batch-files which are text-files with simple commands. In spite of these promising features, Maxent models seem to have a drawback: the models may fail to make general predictions (Peterson *et al.*, 2007).

Maxent creates HSMs by combining presence-only data with environmental layers using a machine-learning approach known as maximum entropy (i.e. that is closest to uniform). Maximum entropy estimates a species' ecological niche by finding a probability distribution which is based on a distribution of maximum entropy under the constraint that the expected value of each environmental variable under this estimated distribution matches its empirical mean (Phillips *et al.*, 2006). This method is equivalent to finding the maximum-likelihood distribution of a species (Phillips *et al.*, 2004). The resulting probability distribution reflects the suitability of the environment for the species of interest. The model evaluates the suitability of each raster cell as a function of the environmental variables at that cell.

We used standard settings of Maxent and a logistic output, with suitability values ranging from 0 (unsuitable habitat) to 1 (optimal habitat) (Phillips and Dudík, 2008). Using standard settings, and thus auto feature selection, implicates that Maxent will automatically add modelling features with increasing number of samples in the training set: below 10 samples only linear functions are used; between 10 and 14 samples quadratic features are added; between 15 and 79 samples hinge features are added and above 79 samples product and threshold features are allowed.

Validation of the models

Whenever data is supplied in the correct format, Maxent will create a habitat suitability model. The question however is whether this model meets all the quality conditions and if

the model output is not influenced by overfitting, preferential sampling and spatial autocorrelation.

Models are qualified using quality parameters. The most commonly used measure is the area under the curve (AUC). It is a threshold independent measure of overall accuracy of the model. It measures the probability that the model will assign a higher probability of occurrence to the observed presences (Bonn and Schröder, 2001). The values of the AUC vary from 0.5 (model not different from random) to 1.0 (perfect accuracy). However, in presence-only modelling the upper limit is always smaller than 1 (Wiley *et al.*, 2003). If the species' distribution covers a fraction a of the pixels, then the maximum achievable AUC is $1 - a/2$. Unfortunately, a is not known, so it is impossible to know how close to optimal a given AUC value is (Phillips *et al.*, 2006).

The AUC is the most commonly used performance parameter. We screened 53 articles where Maxent was used for habitat suitability modelling; in 31 of them the AUC-value was the only quality parameter. Most of these 31 articles mentioned the use of a test set, however for some publications it was not clear if the data was split in a training set and a test set. If no test set is used, this may result in unrealistic high AUC values, because the performance parameter is calculated on the same data that was used to build the model and not on an independent dataset. These 53 articles use fixed thresholds for the AUC to delineate good models. Depending on the source models with an AUC higher than 0.6 (Parisien and Moritz, 2009), 0.7 (Cordellier and Pfenninger, 2009), 0.75 (Elith *et al.*, 2006; Suarez-Seoane *et al.*, 2008; Stachura-Skierczynska *et al.*, 2009), or 0.85 (Brown *et al.*, 2008) are considered to be more informative than random or as good models. Araújo and Guisan (2006) defined a rough guide for classifying model accuracy: 0.6-0.7 poor, 0.7-0.8, average, 0.8-0.9 good and 0.9-1 excellent. Fifteen articles combined the AUC with other parameters and methods to test for significance, such as the test gain (Riordan and Rundel, 2009), null models (Raes and ter Steege, 2007; Ficetola *et al.*, 2009), or with threshold dependent accuracy parameters such as the Kappa statistic (Echarri *et al.*, 2009) or other methods.

Null models

In this study the significance of the models was tested by the use of null models as described by Raes and ter Steege (2007). The general idea behind the null model approach is to create random 'imaginary' species by selecting random spots where the species has been 'observed'. This can be done in two ways: (1) by selecting random points from the entire map area or (2) by selecting points from the stations where nematodes were effectively sampled (Fig. 5.1). The first method will yield random models as if the complete area has been sampled. However, scientists tend to visit some areas more frequently, resulting in collection bias (i.e. preferential sampling). The influence of the collection bias on the accuracy of the HSM depends largely on the range of the values of the environmental variables covered by the stations, known as environmental bias (Kadmon *et al.*, 2004). If sampling is environmentally biased, a HSM is more likely to deviate significantly from a null

model that does not include the bias (Raes and ter Steege, 2007). Thus, if the locations of the 'random species' are restricted to the biased sampling stations, these models are more likely to be significantly different from random. Thus, the second strategy can reveal collection bias or preferential sampling in the dataset. This is important since Maxent predictions are vulnerable to spatial biases in input data (Peterson *et al.*, 2007). The number of observations in the dataset may also influence the AUC value of the model. Therefore 500 null models were calculated for 20 different numbers of observations (Table 5.2). For each group of 500 null models the average AUC and the 95% confidence interval (CI) are calculated. The AUC of each 'real' species model is then compared with the 95% CI of the null models; if the AUC of the real species model is higher than the 95% quantile value, this model is significantly different from random.

Overfitting generally occurs when a model is excessively complex, such as having too many degrees of freedom in relation to the amount of data available. An overfitted model will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data. This predictive performance can be derived from the AUC of the independent test set. In case of overfitting the AUC value of the test set will be significantly lower than the AUC value of the training set. We applied a fivefold cross-validation; the data is split in 5 equal parts (±1 data point) and every data point is assigned once to each of the 5 sets. Five models are then created where each set is used once as a test set and the remaining four fifth of the data is used as training data. Overfitting will decrease the average AUC of the test set while preferential sampling and spatial autocorrelation will have a positive effect on the AUC of the test set. This method allows thus to differentiate between overfitting and preferential sampling because when no cross-validation is applied preferential sampling and spatial autocorrelation will still increase the AUC. But in addition overfitting of the training set (which is the only set used) will also have a positive effect on the AUC-value, because the AUC value is derived from the values of the training set which are estimated too optimistically.

Aside preferential sampling and overfitting, spatial autocorrelation may interfere with the modelling process as well. Species observations may be clustered around certain stations. This may inflate validation statistics by including localities that are not spatially independent (Pearson *et al.*, 2007). In order to check whether this effect is present in the data, we selected the data in each set in such a way that all the data points in the test set are at least 5 or 10 km apart from all the data points in the training set. (Pearson *et al.*, 2007; Murray-Smith *et al.*, 2009).

In total 220 000 random models were created (Table 5.2). The 0.95 quantile values of these random models are then used to delineate the random models from the non-random models of the real species.

		Number of random models
No cross-validation	Complete area	20 x 1 x 500
	Stations	20 x 1 x 500
Cross-validation	Complete area	22 x 5 x 500
	Stations	22 x 5 x 500
	Autocorrelation 5 km	18 x 5 x 500
	Autocorrelation 10 km	18 x 5 x 500

Table 5.2. Number of random models = (subdivisions of the number of data points in the training sets) × (number of models: five in case of cross-validation, one if no cross-validation is applied) × (number of null models).

Species AUC

To delineate random from non-random species models, the AUC values of the real species models are compared with the 0.95 quantile values obtained from the null models. As four modelling techniques were applied for the null models, we followed the same strategy for the species data in order to allow for a valid comparison. This implied modelling (1) without cross-validation (i.e. all the observations where used to create the model); (2) using a fivefold cross-validation; (3) using a fivefold cross-validation, with the data in the test set at least 5 km apart from the data in the training set and (4) a fivefold cross-validation, with the data in the test set at least 10 km apart from the data in the training set. For the latter two techniques the data division algorithm needed to be changed since it was not always feasible to divide all the data in 5 equal parts in such a way that all the points in the five sets are 5 or 10 km apart from all the other points in the other sets. Thus, the data division algorithm was adapted in order to meet two conditions: (1) the number of data in each set is maximised and each set contained more or less the same number of data (±1 data point) and (2) all points in each set are at least 5 or 10 km apart from the data points in the other sets.

Furthermore, it is interesting to assess why certain species models are significantly different from random, while this is not the case for other models. It has been noted before that specialist species, which have specific habitat requirements, tend to have higher AUC values, while generalists have lower AUC values (Elith *et al.*, 2006; Raes and ter Steege, 2007; Lobo *et al.*, 2008; Wollan *et al.*, 2008). Generalists show no specific niche preference and are expected to appear across the complete study area. Therefore we calculated the correlation between the AUC of the species models and four parameters indicating the generalistic occurrence of a certain species: (1) the number of times a species is found in different stations; (2) the niche breadth; (3) the area occupied by the species and (4) the average distance between the stations where the species is found.

The niche breadth of a species was calculated as the mean Euclidean distance of the environmental variables between the stations where the species is found:

$$ED_i = 2.\frac{\sum_{k=1}^{S_i-1}\sum_{l=k+1}^{S_i}\frac{\sqrt{\sum_{j=1}^{N}(x_{jk}-x_{jl})^2}}{N}}{S_i.(S_i-1)} \qquad \text{(Eq. 5.1)}$$

All environmental variables were standardised to mean 0 and standard deviation one. The variable x_{jk} is the standardised value of the environmental variable j at station k where the species is found. N is the total number of environmental variables in the dataset and S_i is the number of stations where the species is found.

The area occupied by the species is estimated by calculating the area included by the straight lines connecting the extreme points of the stations where the species is found.

RESULTS

The results of the six randomisation techniques are summarised in Fig. 5.2 and 5.3. Continuous lines for each of the six techniques are created by interpolation.

Average of randomisations

The average AUC values of the null models derived from the total area are smaller than the AUC values obtained for the null models selected from the actual sampling stations, both for cross-validation and non-cross-validation approaches (Fig. 5.2A and B).

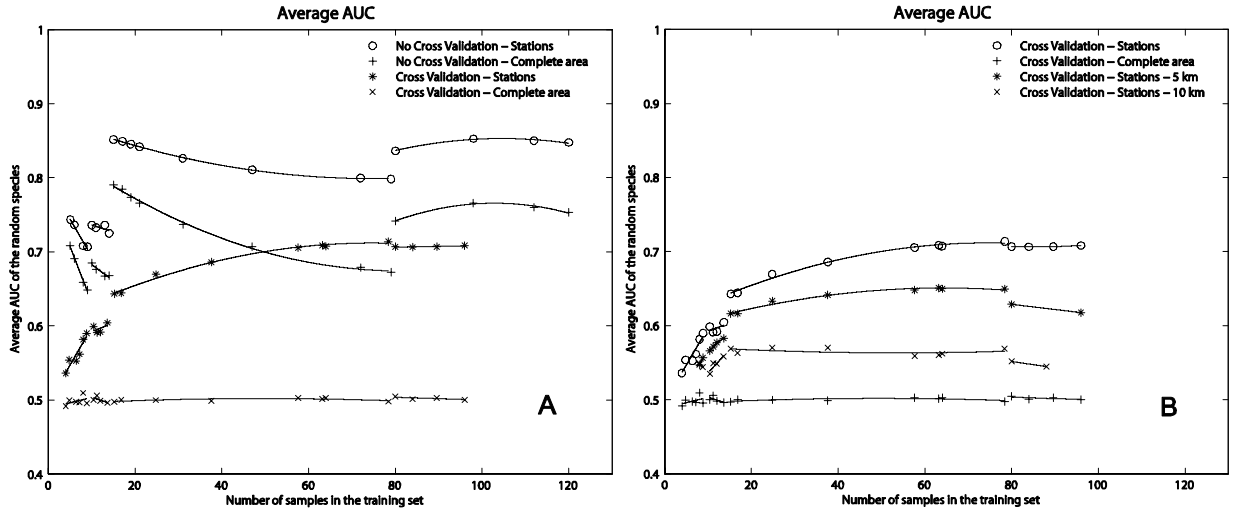

Fig. 5.2. Average AUC of the random models: (A) random samples are selected from the total area or from the sampled stations, both with and without cross-validation and (B) random samples are selected without any restrictions to the sampling distance between the different subsets and with at least 5 km or at 10 km distance between test and training sets.

When no cross-validation is applied the average AUC of the training sets are used since no test sets were created. In this case an increasing number of sampling locations for a set of Maxent modelling features leads to a decrease in the average AUC. However, when all the

features are applied (i.e. when the number of sampling stations > 79) the AUC value stabilises to 0.75 when considering the total area, and to 0.85 when only the sampling stations are used. Adding a modelling feature always results in a strong increase in average AUC.

A completely different pattern is observed when cross-validation is applied. The average AUC is approximately 0.5 when the random samples are selected across the entire area. This value is independent of the number of observations and the added features. This shows that in this case the 'random species' models are truly random. However, when the random observations are restricted to the sampling stations, the average AUC starts off around 0.55 for 5 sampling spots and gradually increases with increasing data points. The curve levels off to an average AUC value of 0.7; thus the test sets of the null models already yields an average AUC of 0.7.

When cross-validation is applied, the addition of features has only a limited effect on the AUC of the test sets. Only in case hinge features are added to the model (i.e. between 14 and 15 samples), a small increase is clear.

Since preferential sampling is clearly present, the effect of spatial autocorrelation was only tested on random observations selected from actual sampling stations. When stations are sampled in such a way that the stations in the test set are at least 5 or 10 km apart from the stations in the training sets, a decrease in the average AUC is clear. This decrease is stronger for datasets with a distance between the data points of at least 10 km.

95% CI of randomisations

When selecting a species model, it is essential to know which model is significantly different from random. Therefore a one-sided 95% confidence interval is constructed to delineate random from non-random species models. Fig. 5.3A and B shows the 0.95 quantile values of the random models. Continuous lines are created by interpolation.

As for the average AUC, the effect of preferential sampling on the 95% CI is clear. The AUC-values of the 'random species' models selected from the sample stations are clearly higher than those selected from the complete area, both for the cross-validation and non-cross-validation approaches.

If no cross-validation is applied there is always a jump to higher AUC values whenever a feature is added. This increase can be considerable: when hinge features are added (between 14 and 15 observations), the AUC-value of the 95% CI jumps from 0.76 to 0.87 when the whole area is considered. These jumps are much smaller and have nearly disappeared when cross-validation is applied. The only observable jump to higher values is in case the observations are chosen from the actual sampling stations and hinge features are added (between 14 and 15 observations).

Without cross-validation the curves of the average AUC and those of the 95% CI are quite similarly shaped. The pattern of the 95% CI-curves is clearly different from that of the

average AUC when cross-validation is applied: the average AUC of the null models is constant or increasing with increasing number of observations in the training set while the opposite is true for the 95% CI. Hence, there is a very high error rate at small sample sizes.

The influence of spatial autocorrelation on the 95% CI is also clear: the AUC decreases with increasing distance between the stations in the training and the test set.

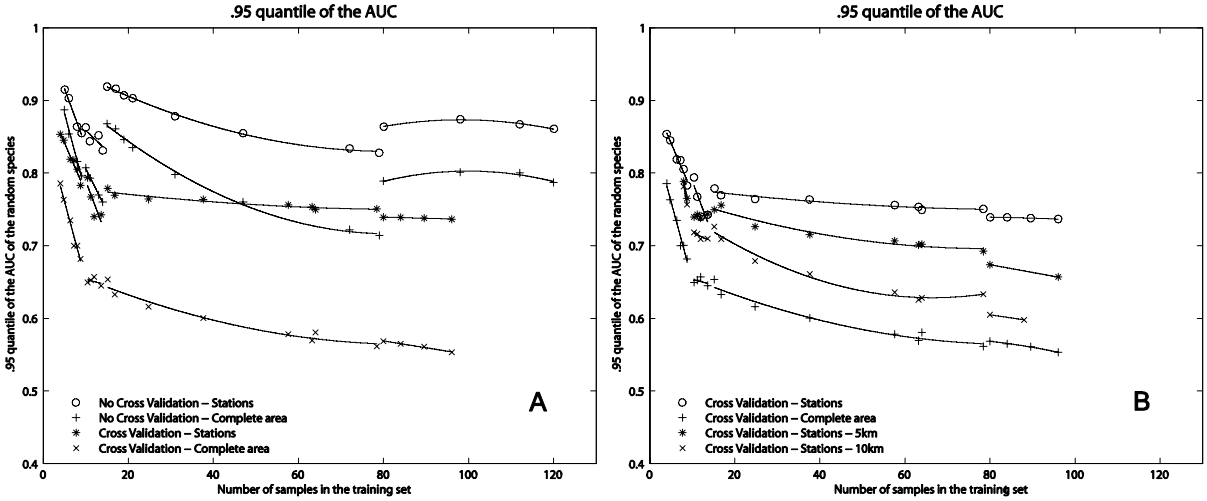

Fig. 5.3. 95% quantile of the random models: (A) random samples are selected from the total area or from the sampled stations, both with and without cross-validation and (B) random samples are selected without any restrictions to the sampling distance between the different subsets and with at least 5 km or at 10 km distance between test and training sets.

Selecting non-random species models

The boundary between random and non-random models is defined by the 95% CI of the AUCs of the random models (Fig. 5.2). The AUCs of the real species models are plotted against these borders (Fig. 5.4 and 5.5). Every test which has been applied on the null models was also applied to the real species data. Thus, four tests have been run on the real species data: with (Fig. 5.4B) and without cross-validation (Fig. 5.4A) and two spatial autocorrelation tests (Fig. 5.5A and B).

When the entire geographical area is sampled and no cross-validation is applied, we found 186 species models (83%) with an AUC higher than the corresponding 0.95 CI. Hence, these models are considered to be significantly different from random. With cross-validation this number even increases to 188. If only the sampling stations are considered these numbers decrease to 126 (no cross-validation) and 122 (cross-validation) (Table 5.3). Notwithstanding

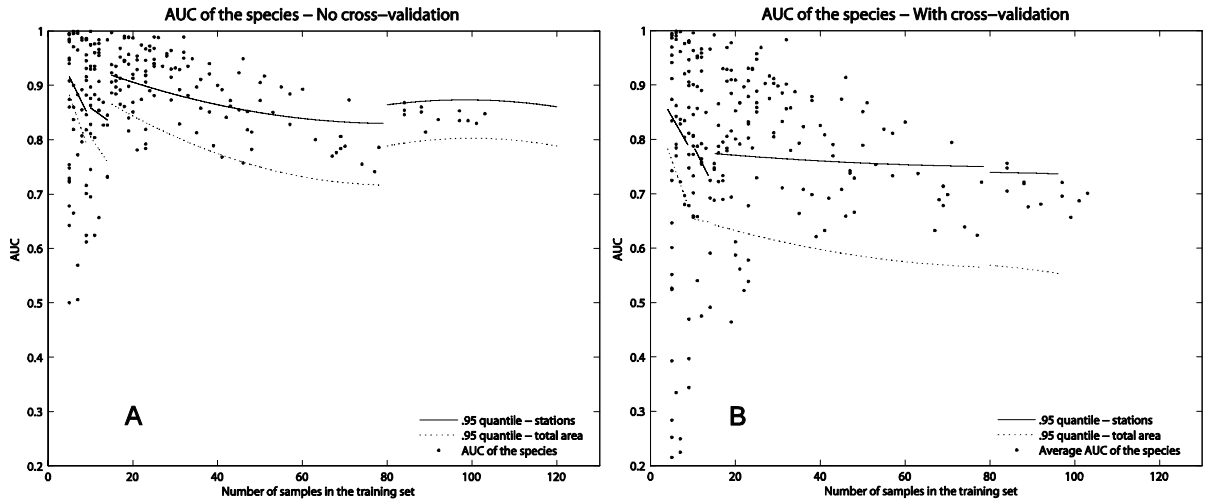

Fig. 5.4. (A and B) Fitted curves of the 95% quantile CI of the null models sampled from the stations and sampled from the total area, with (B) and without (A) cross-validation. AUC values of the species models with (B) and without (A) cross-validation are plotted against these fitted lines (•). Dotted lines are the fitted curves for null models sampled from the total area, full lines result from the null models sampled from the environmentally biased sampling stations.

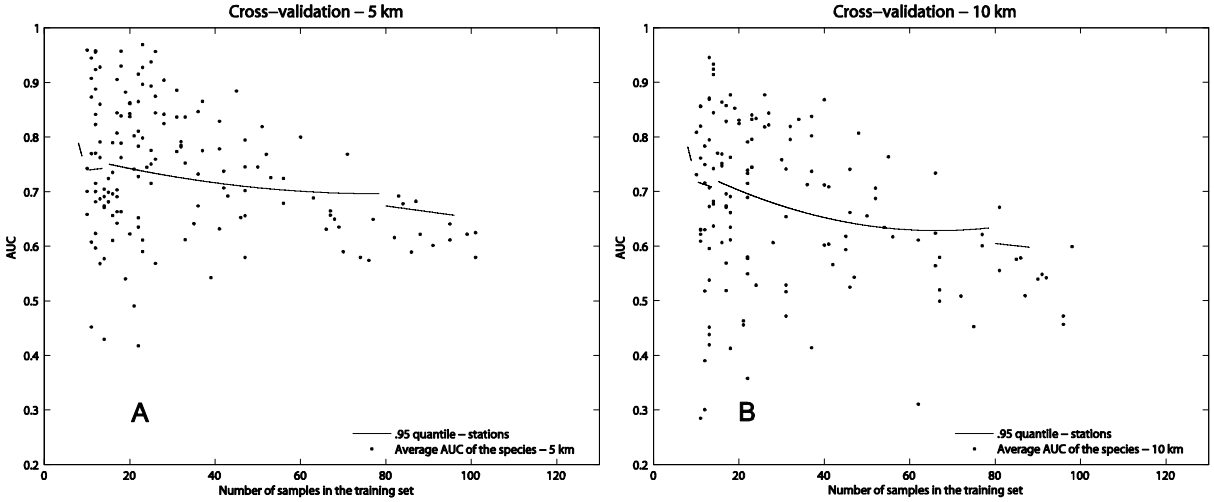

Fig. 5.5. (A and B) Fitted curves of the 95% quantile CI of the null models sampled from the environmentally biased sampling stations with cross-validation and test and training set at least 5 km (A) or 10 km (B) apart. AUC values of the species models with test and training set 5 km (A) or 10 km (B) apart (•).

the fact that with or without cross-validation almost the same number of species models are considered to be significantly different from random, these species are not the same: 111 species pass both tests. From the 15 species which are uniquely selected by cross-validation, 9 models changed from a more complex model without application of cross-validation to a simpler model when cross-validation was applied. This is due to the fact that the number of samples in the training set equals the number of observations of that species. When applying cross-validation, the number of samples in the training set is one fifth smaller,

because one fifth of the data is used for the test set. Since the complexity of feature interactions changes at defined thresholds this explains why the algorithm can shift to simpler interactions.

The Spearman rank correlations between the AUC of the species models and the parameters indicating that a species is a generalist can be found in Table 5.4. All factors show a significant ($p < 0.05$) negative correlation with the AUC, indicating that significant models are not easily created for generalist species. The strongest negative correlation is found between the AUC and the average distance between the stations.

When spatial autocorrelation is considered it seems that the 5 and 10 km subsets could only be created for 150 and 137 species, respectively. Of these species models 76 and 63 pass the 5 km and 10 km test, respectively. Only 54 species models pass all the tests (Table 5.3).

	Minimum distance between cross-validation sets (km)	Number of species models with an AUC higher than the 0.95 CI of the null models		Total number of species analysed
		Total area	Stations	
No Cross-validation	-	186	126	223
Cross-validation	0	188	122	223
Passing both tests		180	111	
CV - 5 km	5	-	76	150
CV - 10 km	10	-	63	137
Passing all tests			54	

Table 5.3. Number of species passing the different tests: preferential sampling, cross-validation and spatial autocorrelation (5 and 10 km).

	Number of observations	Niche breadth	Area occupied by the species	Mean distance between stations of the species
No cross-validation	-0.15	-0.3	-0.6	-0.8
Cross-validation	-0.14	-0.28	-0.62	-0.77

Table 5.4. Spearman rank correlation between the AUC of the species models and parameters indicating a species is a generalist: number of observations, niche breadth, area occupied by the species and average distance between the stations where the species is found.

Spatial autocorrelation not only causes a decrease in interpolated 95% quantile curves, but also results in lower AUC values. The decrease for the species models is even a little stronger than for the curves which are based on the null models. For the 5 km subsets the AUC curves lower on average 0.031 for the null models while the AUC of the species models decrease 0.046. For the 10 km subsets the decrease of the curves is 0.071 for the null models, while for the species models it reaches on average 0.110. A Wilcoxon rank test pointed out that in both cases the decrease for the species models is significantly larger than the decrease in the

0.95 CI. It is thus clear that spatial autocorrelation does indeed inflate the AUC-values of the species models.

DISCUSSION

Three important modelling issues are addressed with this null model approach: preferential sampling, spatial autocorrelation and overfitting. Preferential sampling and spatial autocorrelation are issues linked to the database while overfitting can be attributed to the modelling algorithm.

Average of randomisations

In the ideal scenario, sampling intensity should be equally divided among all sampling stations within a geographical area. In reality this is rarely the case. Preferential sampling is clearly present in our dataset as well. The average AUC of the null models already reaches values around 0.85. Even with cross-validation, average test set values of about 0.7 are not unusual. These random models would be classified as different from random or even as good models according to several sources which have defined a fixed threshold to delineate good from poor models (Cordellier and Pfenninger, 2009; Parisien and Moritz, 2009). This clearly indicates that using a fixed threshold to delineate good models is precarious since most databases are subject to preferential sampling.

If no cross-validation is applied, the average AUC of the null models selected from the complete area is high. Spatial autocorrelation and overfitting may attribute to these high AUC values. Cross-validation helps in differentiating between both effects: spatial autocorrelation leads to high values in the test set, while overfitting will cause lower AUC values. In our case overfitting seems to be strongly present because the average AUC of the test set is much lower. Cross-validation thus clearly reveals overfitting. However, since one fifth of the data is used for testing, a disadvantage of cross-validation is that less of the available information can be used to construct the model.

If no cross-validation is applied there are strong jumps whenever a feature is added to the algorithm. These increases in AUC can result from overfitting or from an improvement in the model owing to the extra feature. Cross-validation again helps in distinguishing between these two phenomena: the jump to higher AUC-values will disappear in the case of overfitting because the test set will not yield better results. It is clear from Fig. 5.2A that these jumps can be mainly attributed to overfitting. Only the addition of hinge features seems to improve the AUC of the test set. Thus, adding hinge features helps explaining the variation in the data. However, in this case it is peculiar, because the samples are randomly picked from the sample stations and this improvement must thus be attributed to preferential sampling.

The influence of preferential sampling is stronger with increasing number of observations in the training set when cross-validation is applied. This is caused by the increasing chance of

incorporating samples from the preferentially sampled area in both the training and test set, with increasing sample numbers.

Random models with an average AUC of 0.5 are only observed in the case of cross-validation combined with random sampling across the whole region. An increase in the average AUC is observed when only the sampling stations are considered during modelling, which can be attributed to preferential sampling or to spatial autocorrelation of the samples. Both aspects are not clear when only a few stations are sampled, but with an increasing number of samples these effects become more obvious. Autocorrelation is a difficult topic to tackle, because it is difficult to differentiate between spatial autocorrelation and regionally restricted species with strong environmental preferences. Spatially separating test and training set clearly lowers the AUC of the test set, meaning that the unseen test data is harder to predict. If no spatial division is made for the test and the training set the AUC of the test set is considerably higher. Thus spatial autocorrelation clearly influence the results of the models.

As such, we showed clearly that combination of preferential sampling, spatial autocorrelation and overfitting lead to inflated AUC values of 0.85 for a random model while on average it should have an AUC of 0.5.

95% Quantile of the randomisations

The 0.95 quantile curves are used to distinguish random from non-random species models. It is clear that without cross-validation models with AUC-values as high as 0.9 are not necessarily different from random. With cross-validation and at low sample sizes AUC values of 0.85 are not unusual. With increasing sample numbers this value decreases to about 0.75. It is thus clear that the predefined thresholds are not applicable to this dataset.

Although the four curves look quite similar, it is clear that the curves obtained after cross-validation are again less sensitive to the addition of a feature.

The test for spatial autocorrelation (Fig. 5.3B) shows that the AUC of the null models decreases with increasing distance between stations in the test and training set. This is not surprising because the chance of sampling a different environment increases with increasing distance between the stations, which makes it hard to predict the values of the test set.

Selecting non-random species models

The 0.95 quantile curves allow for significance testing of HSMs. Species models performing better then random reflect species with specific niche requirements that can be relatively easy predicted. On the other hand, the reason why species models are performing worse than random may be attributed to different causes: (1) the species are generalists and have no specific environmental requirements; (2) the environmental variable explaining the distribution of the species is not available; (3) the distribution of the species is not well estimated because of a sampling bias. The generalist theory is further supported by the

strong negative correlation between the AUC and the average distance between two sites where the species is found. This strong correlation indicates that the species which are not confined to a limited area are hard to predict. The negative correlation between the niche breadth (based on the environmental variables and not on the location in space) and the AUC is less strong but still significant, thus the variation in environmental space can also partly explain why some species are harder to predict than others.

If the sampling locations are environmentally biased this may lead to HSMs predicting an underestimation of the true geographical range of the species (Raes and ter Steege, 2007). In our case preferential sampling has clearly the strongest effect on the selection of the species models.

Subsets meeting the 5 km distance criterion could be created only for 150 out of 223 species. This indicates that the 73 other species are strongly concentrated in space, making it impossible to find five samples with a minimum distance of 5 km from each other. However, does this mean that the models of these 73 species are inadequate? This would suggest that it would be impossible to correctly predict the distribution of species restricted to a small area. We believe this is not necessarily the case; however, the AUC-values of these models should be treated with caution. It has been shown that spatial autocorrelation may represent a problem for species' distribution models. Significance values of the models may be severely inflated (Segurado *et al.*, 2006) because the test and training set are not entirely independent. Also the choice of the environmental variables by the model is questionable. Indeed, all environmental variables show spatial autocorrelation. Therefore, all the environmental variables have more or less the same value within this area. Thus, the selection of the environmental variables explaining the distribution of the species may be arbitrary.

This methodology allows distinguishing between random and non-random species models. However, when these models are used for management purposes it is important that the models are able to predict unseen data correctly and have a good predictive performance. Although this approach can reveal overfitting, it is not solving the problem. An advantage of Maxent is that it is able to counteract overfitting by choosing the regularisation setting. We used the default value of 1 (Phillips *et al.*, 2006). It is clear from Fig. 5.4 that overfitting is still present. Overfitting can be further dealt with by setting a different regularisation multiplier, by feature selection or by selecting fewer environmental factors. In this way the reduced model will have a better predictive performance with unseen data. In our case the final models were selected by backwards and forwards selection of the environmental factors (Addendum 3).

In addition to the research of Raes and ter Steege (2007), we also investigated the influence of spatial autocorrelation. Spatial autocorrelation also attributes to the inflated AUC-values.

The modelling issues which are clearly present in this historical database are not necessarily present in every database. Sampling campaigns which are set up according to the statistical principle of random and independent sampling, will not suffer from preferential sampling

and spatial autocorrelation. However, to assure that samples are truly independent, the extent of the range of spatial autocorrelation should be known before sampling starts, which is often not the case. The issue of overfitting is a modelling issue and should always be addressed during modelling.

Drawbacks

Despite many interesting features of the methodology described here, there are some drawbacks as well: our approach is labour intensive and not applicable to all datasets. There is a need for a lot of sampling stations where the species has not been detected. This does not necessarily mean that the species is absent in these stations, with inconspicuous species as nematode species absence is never certain. However, these stations where the species is not detected can be interpreted as a station with a low presence probability or as a pseudo-absence, similar to the back-ground data used by the algorithm. In contrast however, these 'pseudo-absences' are not uniformly chosen but restricted to the sampling stations.

Maxent is applicable to specialist species. However, with this technique it is not possible to delineate generalist species models from null models although the model may reflect the true habitat of the generalist. Nevertheless, if the conservation biologist is mainly interested in rare and specialist species this will not be an issue.

CONCLUSIONS

Our results show clearly that the commonly used thresholds (Araújo and Guisan, 2006; Elith *et al.*, 2006; Brown *et al.*, 2008; Suarez-Seoane *et al.*, 2008; Cordellier and Pfenninger, 2009; Parisien and Moritz, 2009; Stachura-Skierczynska *et al.*, 2009) are not readily applicable to all datasets and should be treated with caution. Many aspects may influence and inflate the final AUC-value of a HSM. Therefore, a thorough examination of the dataset is necessary: is there sample bias and thus preferential sampling in the dataset? Can spatial autocorrelation partly explain the high AUC values of the models? Is overfitting present and can it be tackled? These questions are not always addressed, but it is clear that these aspects strongly influence the AUC: inflations of the AUC from 0.5 to 0.9 are possible. For habitat suitability models this is the difference between a random model and a good model!

ACKNOWLEDGEMENTS

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. This research was conducted within the MANUELA framework (http://www.marbef.org/projects/Manuela), which is a Responsive Mode Project undertaken within the MarBEF EU Network of Excellence 'Marine Biodiversity and Ecosystem Functioning' which is funded by the Sustainable Development, Global Change and Ecosystems Programme of the European Community's Sixth Framework Programme (Contract No. GOCE-CT-2003-505446). This research was also supported by the GENT-BOF Project 01GZ0705 Biodiversity and Biogeography of the Sea (BBSea).