CHAPTER 3

PREDICTABILITY OF MARINE NEMATODE BIODIVERSITY

Adapted from: Merckx, B., Goethals, P., Steyaert, M., Vanreusel, A., Vincx, M., Vanaverbeke, J., 2009. Predictability of marine nematode biodiversity. Ecological Modelling 220, 1449-1458.

CHAPTER 3

PREDICTABILITY OF MARINE NEMATODE BIODIVERSITY

ABSTRACT

In this paper, we investigated: (1) the predictability of different aspects of biodiversity, (2) the effect of spatial autocorrelation on the predictability and (3) the environmental variables affecting the biodiversity of free-living benthic marine nematodes on the Belgian Continental Shelf. An extensive historical database of free-living marine nematodes was employed to model different aspects of biodiversity: species richness, evenness, and taxonomic diversity. Artificial neural networks (ANNs), often considered as 'black boxes', were applied as a modelling tool. Three methods were used to reveal these 'black boxes' and to identify the contributions of each environmental variable to the diversity indices. Since spatial autocorrelation is known to introduce bias in spatial analyses, Moran's I was used to test the spatial dependency of the diversity indices and the residuals of the model. The best predictions were made for evenness. Although species richness was quite accurately predicted as well, the residuals indicated a lack of performance of the model. Pure taxonomic diversity shows high spatial variability and is difficult to model. The biodiversity indices show a strong spatial dependency, opposed to the residuals of the models, indicating that the environmental variables explain the spatial variability of the diversity indices adequately. The most important environmental variables structuring evenness are clay and sand fraction, and the minimum annual total suspended matter. Species richness is also affected by the intensity of sand extraction and the amount of gravel of the sea bed.

Keywords:

Biodiversity, marine, Nematoda, spatial autocorrelation, artificial neural networks

INTRODUCTION

As a consequence of the ever increasing anthropogenic pressure on the sea floor, there is a growing need for sustainable management of this vulnerable environment. These management decisions have to be based on sound scientific data concerning the functioning of the environment and the diversity of the benthic organisms. Biodiversity indices are often used to describe areas of high biological interest. Biodiversity, however, is a broad concept covering different aspects of a community, e.g. evenness, taxonomic diversity, and species richness. Species richness is the most commonly used indicator, but it is highly dependent on sampling effort. This is not an issue in datasets collected by a single investigator. However, in large datasets originating from different sources, sampling strategy and effort can vary considerably. Therefore, we focused on indices which are assumed to be independent of sampling effort: estimators for total species richness (Chao, 1984, 1987) and evenness (Chao and Shen, 2003), the expected species richness (Sanders, 1968; Hurlbert, 1971; Simberloff, 1972), and taxonomic diversity indices (Clarke and Warwick, 1998). These diversity indices can exhibit spatial autocorrelation (SA), meaning that nearby observations are more similar than observations farther away (Odland, 1988; Legendre, 1993). Although, spatial autocorrelation can be an important source of bias in spatial analyses (Segurado *et al.*, 2006), it is often ignored in ecological studies (Dormann, 2007). If SA remains in the residuals of the model, it may even invert the observed pattern of an environmental variable (Kühn, 2007). Although SA should always be investigated, it does not necessarily generate bias, and should be considered a tool to investigate the factors influencing richness on different spatial scales (Diniz-Filho *et al.*, 2003).

Studies of the freshwater environment employed artificial neural networks (ANNs) to predict the occurrence of macrobenthic invertebrates (Dedecker *et al.*, 2004) and diversity measures (Park *et al.*, 2003). ANNs are a data driven modelling technique which received increased attention in ecological sciences as a powerful, flexible tool for uncovering complex patterns in data (Park *et al.*, 2005). These models can simulate any continuous mathematical function and are therefore more appropriate to describe complex ecological functions than linear models. In spite of their appealing characteristics, their exploratory value is often criticised, being coined a 'black box' approach (Lek *et al.*, 1996a) since the contribution of the input variables to the output is difficult to disentangle from the network. Several methods have been proposed to eliminate this problem, three have been applied herein: the Perturb (Yao *et al.*, 1998; Scardi and Harding, 1999; Gevrey *et al.*, 2003), the Profile (Lek *et al.*, 1995, 1996a, b; Gevrey *et al.*, 2003) and a Modified Profile algorithm.

To our knowledge, similar modelling efforts on marine free-living nematodes (part of the meiobenthos) have not been attempted yet. This is surprising since free-living nematodes represent the highest metazoan diversity in many benthic environments in terms of species numbers (Heip *et al.*, 1985): more than 50 species are commonly found in a single 10 cm^2 core. Owing to their interstitial life style, biogeochemical properties of the sediment have a

strong influence on the diversity and the composition of nematode assemblages (Heip *et al.*, 1985; Steyaert *et al.*, 1999). Therefore, nematode-biodiversity studies are appropriate to assess environmental impact in the marine benthic environment (Heip *et al.*, 1985; Kennedy and Jacoby, 1999; Boyd *et al.*, 2000; Schratzberger *et al.*, 2000a). Moreover, nematode communities seem to be resilient and their restoration occurs easily after temporal, low impact disturbances (Kennedy and Jacoby, 1999; Schratzberger *et al.*, 2002), making them a pertinent community to model based on long term environmental data.

In this paper, nematode biodiversity on the Belgian Continental Shelf (BCS) was modelled with a wide range of environmental variables, and the following issues were addressed: (1) the predictability of different aspects of biodiversity, (2) the effect of spatial autocorrelation on the predictability and (3) the environmental variables affecting these biodiversity indices for free-living marine nematodes.

METHODS

Study area

The Belgian Continental Shelf (BCS) is situated in the southern part of the North Sea (Fig. 3.1). The total surface area is 3600 km^2 (approximately 0.5% of the total area of the North Sea), and reaches 42 m depth. The seabed is a heterogeneous environment characterised by shallow sandbanks and a broad spectrum of sediments, ranging from clay to coarse sands (Lanckneus *et al.*, 2002).

Biological data

Within the EU Network of Excellence MarBEF, MANUELA is a Responsive Mode Project focusing on the meiobenthos (metazoans passing a sieve of 1 mm and retained on a 38 µm sieve). A central MANUELA database was compiled comprising the available data on meiobenthos on a broad European scale (Vandepitte *et al.*, 2009). We restricted the analyses to the BCS, since extensive environmental data were available for this region. The final dataset consisted of 209 samples belonging to 75 different stations on the BCS (Fig. 3.1), collated from nine different datasets. This data includes information on 29 783 nematodes identified to species level and collected in the period 1972-2004.

Environmental predictors

Some of the environmental variables were measured during sampling and could be retrieved from the MANUELA database. However, for most of the environmental predictors no data was readily available and these values were retrieved from area covering maps of the Belgian Continental Shelf (Table 3.1).

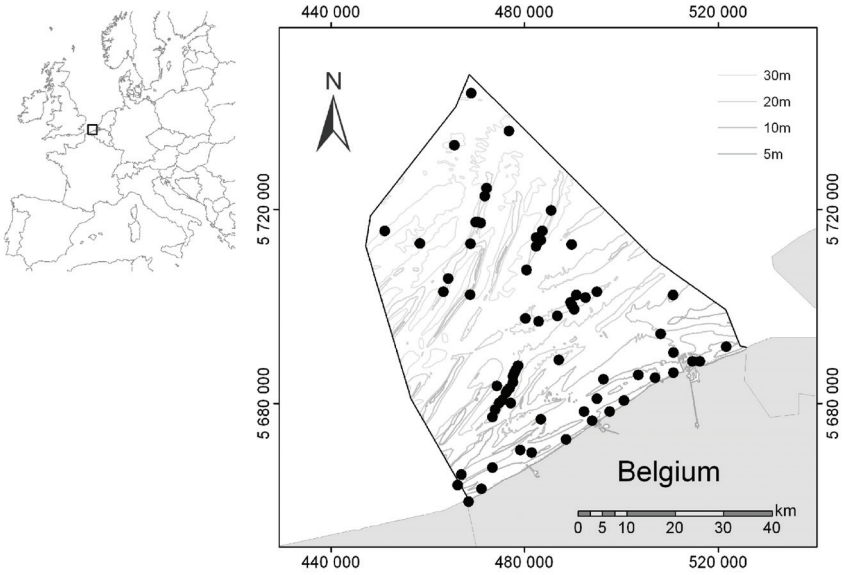

Fig. 3.1. Sampling stations on the Belgian Continental Shelf (•) (UTM31N-WGS84 coordinates).

A map of the intensity of sand extraction, representing the number of extractions per year at a certain raster cell, was constructed with data collected by the Fund for Sand Extraction during the years 1996-2005 (data courtesy of the Federal Institute of Economics). The construction of one single map for the whole period is acceptable, since comparison of the recent data with older publications (Rzonzef, 1993) showed that the regions where sand extraction occurred remained unchanged.

Biochemical data and current properties data were supplied by the Management Unit of the North Sea Mathematical Models and the Scheldt estuary (MUMM). Chlorophyll *a* and total suspended matter (TSM) data were collected with the MERIS spectrometer on board of the ENVISAT satellite of the European Space Agency. Eighty chlorophyll *a* maps from the period 2003-2005 were reduced to only three maps with the minimum, maximum and average values. Likewise, 90 TSM maps (2002-2005) and 27 salinity maps (1996-2002) were converted into three maps.

Oceanographic and sedimentological data were supplied by the Renard Centre of Marine Geology (RCMG) of Ghent University. The Bathymetric Position Index indicates whether a raster point is situated on a peak, in a gully or on a plain.

The year and date of sampling were used as temporal variables. The variable representing the time of year should have the same value at the end of December and at the start of January. Therefore, we transformed the date into a variable showing a standard normal distribution with a maximum situated around the 1st of August, the warmest period of the year.

Variable type		Variable	Abbreviation	Unit	Min	Max	Moran's I (p<0.001)
Anthropogenic	*	Intensity of sand extraction	sand_extr	#/year	0	14890	0.37
Biochemical		Average total suspended matter	TSM_mean	$g.m^{-3}$	2.1	24.4	0.97
		Maximum total suspended matter	TSM_max	$g.m^{-3}$	3.9	43.7	0.9
	*	Minimum total suspended matter	TSM_min	$g.m^{-3}$	0.58	9.95	0.82
	*	Average chlorophyll content	chl_mean	$mg.m^{-3}$	2.3	17.6	0.7
		Maximum chlorophyll content	chl_max	$mg.m^{-3}$	6.6	35.2	0.82
	*	Minimum chlorophyll content	chl_min	$mg.m^{-3}$	0.04	4.27	0.24
		Average salinity	sal_mean		30.2	35.2	0.97
	*	Maximum salinity	sal_max		31.5	35.5	0.9
		Minimum salinity	sal_min		29.5	34.5	0.92
Current	*	Minimum bottom shear stress	Bstri	$N.m^{-2}$	0	0.105	0.82
		Mean bottom shear stress	Bstrm	$N.m^{-2}$	0.008	1.08	0.47
		Maximum bottom shear stress	Bstrx	$N.m^{-2}$	0.07	8.33	0.21
		Size of the residual currents	Mcur	$m.s^{-1}$	0.002	0.077	0.39
	*	Maximum depth-averaged current velocity	Mmax	$m.s^{-1}$	0.12	1.17	0.18
		Magnitude of the residual transports	Mtra	$m.s^{-1}$	0.002	0.138	0.49
	*	Residual currents	Rcur	$m.s^{-1}$	-0.023	0.07	0.33
		Residual transports	Rtra	$m.s^{-1}$	-0.029	0.121	0.5
	*	Tidal amplitude	Tampl	m	2.79	5.41	0.95
		Maximum current velocity at the bottom layer	Umax	$m.s^{-1}$	-0.66	1.11	0.25
	*	Average current velocity at the bottom layer	Umea	$m.s^{-1}$	0.04	0.62	0.54
Topographic		Water depth	depth	m	2.2	41.9	0.64
	*	Slope of the sea bottom	slope	°	0.03	2.89	0.51
	*	Bathymetric Position Index (1600 m range)	bpi_1_20	-	-495	206	0.48
	*	Bathymetric Position Index (240 m range)	bpi_1_3	-	-316	183	0.27
		Rugosity of the bottom	rugosity	$m^2.m^{-2}$	1	1.0014	0.28
		Orientation of the slope of the bottom	aspect	°	34	354	0.16
Sediment	*	Median grain size	d50	µm	38	654	0.56
	*	Gravel content	gravel	weight%	0	34	0.24
	*	Sand content (63 µm - 2 mm)	sand	%	4.7	100	0.7
	*	Silt-clay content (0-63 µm)	mud	%	0	95	0.7
Time		Year of collection	year	year	1972	2004	
		Annual cycle (maximum on August, 1st)	date	-	0.002	0.19	

*Table 3.1. Abiotic factors used in the model with their minimum and maximum values and Moran's I. * indicates if this variable contributed more than 5% in at least one model.*

Biodiversity indices

Species richness

Species richness S, based only on occurrence data, is affected by the sampling effort. Therefore, this index was compared with Chao's moment estimators of species richness, which are less prone to sampling effort. Chao (1984, 1987) developed different moment estimators of the lower bound of species richness, using the information of the number of species sampled once (f_1) and twice (f_2). Two estimators were used here (Chao, 1984; from Chiarucci *et al.*, 2003):

$$S_{ch1} = S + \frac{f_1(f_1-1)}{2(f_2+1)} \qquad \text{(Eq. 3.1)}$$

$$S_{ch2} = S + \frac{{f_1}^2}{2(f_2+1)} - \frac{f_1 f_2}{2(f_2+1)^2} \qquad \text{(Eq. 3.2)}$$

with S the total number of species observed in the sample. Although these indices are a lower bound rather than an estimate, they have been shown to work well as an estimator (Chao, 1984; Hortal *et al.*, 2006).

Expected species richness

When individuals are independently sampled with similar probability from a small sample the expected species richness, if the sample was of the smaller size n, is (Sanders, 1968; Hurlbert, 1971; Simberloff, 1972):

$$ES(n) = \sum_{i=1}^{S}\left(1 - \frac{\binom{N-x_i}{n}}{\binom{N}{n}}\right) \qquad \text{(Eq. 3.3)}$$

where N is the total number of individuals in the sample, x_i the number of individuals of species i, and n is the number of individuals in the subsample. This index is used for interpolating, not for extrapolating to a larger sample size (Gotelli and Graves, 1996); therefore, the n-value was restricted to 200. The other n-values applied here are 20, 25, 50, 100 and 150.

Taxonomic diversity indices (Clarke and Warwick, 1998)

Taxonomic diversity (Δ) reflects the average taxonomic distance between any two organisms, chosen at random from a sample. The distance can be seen as the path length connecting these two organisms through a phylogenetic tree or a Linnean classification. This index includes aspects of taxonomic relatedness and evenness (Clarke and Warwick, 2001).

$$\Delta = \frac{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S}\omega_{ij}x_i x_j}{N(N-1)/2} \qquad \text{(Eq. 3.4)}$$

where x_i is the abundance of the i^{th} species, N the total number of the individuals in the sample, ω_{ij} the distinctness weight given to the path length linking species i and j in the hierarchical classification and S is the number of species.

Taxonomic distinctness (Δ^*) is the average path length between two randomly chosen but taxonomically different organisms. This value is a measure of pure taxonomic relatedness, although abundances are still used to calculate this index.

$$\Delta^* = \frac{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S}\omega_{ij}x_i x_j}{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S}x_i x_j} \qquad \text{(Eq. 3.5)}$$

When only occurrence data is considered, both Δ and Δ^* converge to the same statistic: the average taxonomic distinctness (Δ^+), which can be seen as the average taxonomic path length between any two randomly chosen species.

$$\Delta^+ = \frac{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S}\omega_{ij}}{S(S-1)/2} \qquad \text{(Eq. 3.6)}$$

Evenness

Shannon's index of diversity, like species richness, also depends on the sampling effort. Nevertheless, this index is included in the analysis to compare its performance with Chao's nonparametric estimation of this index (Shannon, 1948).

$$H' = -\sum_{i=1}^{S} p_i \log_e (p_i) \qquad \text{(Eq. 3.7)}$$

where p_i is the proportion of species i relative to the total number of species.

A nonparametric estimation of Shannon's index of diversity (H_{ch}) was proposed by Chao and Shen (2003). This approach adjusts Shannon's index for unseen species.

$$H_{ch} = -\sum_{i=1}^{S} \frac{\frac{\hat{C}.f_i}{N}\log_e\left(\frac{\hat{C}.f_i}{N}\right)}{1-\left(1-\frac{\hat{C}.f_i}{N}\right)^N}.I(A_i) \qquad \hat{C} = 1 - \frac{f_1}{N} \qquad \text{(Eq. 3.8)}$$

where f_i is the number of species with i individuals in the sample, A_i denotes the event that the i^{th} unit is included in the sample, and $I(A_i)$ is the indicator function ($I(A_i)$ = 1 when A_i is true and $I(A_i)$=0 otherwise).

Model building and selection

The performance of a neural network is improved by implementing consecutive optimisation steps. Firstly, the data needs to be preprocessed; as different variables span different ranges, the data have to be standardised to ensure that all variables receive equal attention during the training. The environmental variables were transformed to mean zero and standard deviation one. This is also a standard procedure when using principal components (see below). The minimum and maximum values of the biodiversity indices were normalised to the interval [−1, 1] (Shi, 2000).

The second step is to design the neural network. Four design criteria for ANNs are distinguished (Walczak and Cerpa, 1999): selection of input variables, design of the number of hidden layers, selection of the number of hidden neurons for each layer, and selection of a learning method. The number of input variables was reduced by a principal component analysis (PCA), and only PCs contributing more than 1% to the variability in the dataset, were retained. In total, 13 PCs were retained explaining 96% of the variability in the dataset. The number of layers in the neural network was restricted to one, since theoretically an ANN with one hidden layer can approximate any function as long as sufficient neurons are used in the hidden layer (Hornik *et al.*, 1989). The optimum number of neurons, the transfer functions, and the learning methods were obtained by comparing the total root mean squared error of the test set for the 10-fold cross-validation. A stratified 10-fold cross-validation was applied to the data, where each environment is represented equally in each of the 10 subsets. In this way, the variation produced by pure random selection of the subsets is reduced (Witten and Frank, 2000). The subsets were created through a fuzzy clustering algorithm (Kaufman and Rousseeuw, 1990; Shahin *et al.*, 2004): once the clusters were created, samples within each cluster were assigned to one of 10 subsets. It is a well known phenomenon that spatially autocorrelated data can inflate the perceived ability of models to make realistic predictions (Segurado *et al.*, 2006); therefore replicates of nearby samples were retained within the same subset. The neural network was trained with 8 out of 10 sets, one set was used as a validation set to prevent overtraining, and one as an independent test set to validate the model.

The computation of a neural network starts with the assignment of randomised weights, which are updated during the optimisation process. Depending on these initial weights, it may be impossible to find the optimal network and a suboptimal solution is selected. Consequently, it is necessary to calculate the network several times, e.g. for the estimation of the number of neurons 500 neural networks per diversity index and per neuron level were calculated. For the final model, after selection of the ANN architecture, 10 out of 500 models for each diversity index were selected to test the stability and the variation between the models (Gevrey *et al.*, 2005).

Model accuracy and precision were assessed by comparing the observed and predicted values of the learning and the test set by means of the Pearson product-moment correlation coefficient, and the concordance correlation coefficient (CCC) (Lin, 1989). The Pearson product-moment correlation coefficient offers only information on the precision of the linear association, while the CCC reports also on the accuracy of the association. The residuals of the models were tested for normality and spatial autocorrelation.

Spatial autocorrelation

The presence and magnitude of spatial dependence in data can be estimated by different statistics. Here, we applied Moran's I (Moran, 1950), a commonly used index ranging from −1 (indicating strong negative spatial autocorrelation (SA) or checkerboard patterns) to +1

(indicating strong positive SA). This index was calculated for both the original response variable and the model residuals, and in both cases for two lag distances, 0-50 m and 50-5000 m. The small lag distance accounts for the SA between replicate samples which are taken within a distance of tens of metres and the larger lag distance assesses the SA for the nearest stations which are mostly located within a lag of 5 km. Significance was tested with a Monte Carlo permutation test (Sawada, 1999). If any spatial structure remains in the residuals, this indicates the existence of at least one lacking variable having a spatially structuring effect on the biodiversity index (Odland, 1988).

Input variables contribution methods

Neural networks are often considered to be 'black boxes', indicating that the contribution of the variables to the output is not easy to disentangle. Accordingly, several input contribution methods have been developed to reveal the importance of the input variables. It is necessary to apply a variety of methods, since other techniques may classify variables differently. In that case, the network is poorly calibrated or the data is difficult to analyse (Gevrey *et al.*, 2003). Since the PCs and not the original environmental variables are the input variables for the neural network, the relation between the original variables and the connection weights is difficult to unravel. Therefore, we applied methods which modify the original environmental variables and excluded those techniques which alter the connection weights.

The contribution of the environmental variables to the model output was assessed in three ways. The first technique, the Perturb method (Yao *et al.*, 1998; Gevrey *et al.*, 2003), estimates the effect of small changes in each environmental variable on the output of the network. The algorithm adjusts the input values of one variable while keeping all the others untouched and records the change of the diversity index. In this way, a classification of the abiotic variables by order of importance is obtained (Yao *et al.*, 1998; Scardi and Harding, 1999). Secondly, we implemented the Profile method (Lek *et al.*, 1995, 1996a, b; Gevrey *et al.*, 2003). The effect of each input variable on the output is observed successively while the other input variables maintain fixed values: their minimum values, first quartile, median, third quartile and maximum. This gives a set of profiles with the variation of the target variable according to the change of each input variable. Finally, we implemented a modified version of the Profile method, which incorporates aspects of the Perturb method. Whereas in the Profile method all the other variables are kept at fixed values, we retained the actual values of the environmental variables. Thus, a more 'natural' environment was created. The selected variable was set at its minimum value, and the diversity index was calculated for all the samples. The average and confidence interval were calculated for these output values. This process was repeated for several values between the minimum and maximum value of the selected environmental variable and subsequently for each environmental variable. Likewise, a set of profiles was created, now with a confidence interval indicating if the

variation of the variable significantly influenced the output. The latter two methods provide information on both the importance and the sign of the environmental variables.

RESULTS

Neural network

The best performing network for all the diversity indices was a network with two neurons in one hidden layer with the hyperbolic tangent and a linear function as transfer functions connecting the input, hidden and output layer.

Spatial autocorrelation and biodiversity indices

Spatial autocorrelation (SA) was present in all biodiversity indices: each biodiversity index displayed a highly significant and positive value of Moran's I ($p \leq 0.001$) (Table 3.2). It was clear that all the models accounted significantly for the SA, since only a fraction of the original SA remained in the residuals. For the replicates (lag 50 m), 65% of the 140 models

	lag 50 m			lag 50-5000m		
Diversity index	Moran's I ($p \leq 0.001$)	Moran's I of the residuals	# of models without SA	Moran's I ($p \leq 0.001$)	Moran's I of the residuals	# of models without SA
S	0.68	0.12 (± 0.03)	3	0.31	-0.009 (± 0.011)	8
S_{ch1}	0.62	0.01 (± 0.03)	9	0.3	0.001 (± 0.006)	10
S_{ch2}	0.62	0.01 (± 0.02)	9	0.3	-0.003 (± 0.007)	9
ES(20)	0.71	0.09 (± 0.02)	5	0.37	-0.004 (± 0.004)	10
ES(25)	0.71	0.14 (± 0.04)	5	0.38	0.004 (± 0.009)	9
ES(50)	0.71	0.05 (± 0.02)	8	0.38	0.002 (± 0.005)	10
ES(100)	0.7	0.07 (± 0.02)	7	0.36	0.001 (± 0.010)	8
ES(150)	0.69	0.08 (± 0.01)	8	0.34	0.000 (± 0.007)	9
ES(200)	0.68	0.06 (± 0.02)	7	0.32	-0.010 (± 0.011)	8
Δ	0.71	0.14 (± 0.03)	4	0.36	0.027 (± 0.007)	6
Δ^*	0.55	-0.03 (± 0.02)	10	0.38	-0.005 (± 0.007)	9
Δ^+	0.26	-0.05 (± 0.02)	8	0.16	-0.006 (± 0.003)	10
H'	0.68	0.13 (± 0.03)	6	0.37	0.018 (± 0.010)	8
H_{ch}	0.7	0.12 (± 0.02)	3	0.39	0.011 (± 0.005)	10

Table 3.2. Spatial autocorrelation of the diversity indices for lag 50 m and 50m-5000m: 1) Moran's I of the diversity index; 2) Average Moran's I of the residuals of the ten selected models (± SE); 3) Number of models with no significant SA in the residuals ($p \geq 0.05$).

showed no significant SA (p > 0.05) and less than 9% displayed highly significant SA. For the longer distance (lag 50-5000m), 89% of the models did not show significant SA and none of the models exhibited highly significant SA.

The results of model performance are presented in Table 3.3. The Pearson product-moment correlation coefficient and the CCC resulted in similar model performances (Fig. 3.2). The number of species could be accurately predicted with an average CCC of 0.79 of the test sets. Likewise, the extrapolated values of the number of species (S_{ch1} and S_{ch2}) showed the same performance. A high CCC-value was found for ES(20) with an average *r* of 0.89; however, by increasing n a decrease in CCC was observed. Ultimately, the r and CCC-values of ES(200) approximated those of S. The taxonomic diversity index Δ was accurately predicted, while the pure taxonomic relatedness Δ^* proved much harder to predict. When only presence/absence was considered (Δ^+), the model accuracy dropped further to 0.45 and 0.41 for *r* and CCC, respectively. The evenness parameters (H' and H_{ch}) produced very high *r*- and CCC-values, comparable to those of ES(20).

A very strong correlation between Moran's I of the biodiversity index and the CCC of the test set is found (*r* = 0.97 for lag 50m and *r* = 0.83 for lag 50-5000m) (Tables 3.2 and 3.3), implying that the performance of the model strongly correlates with the spatial autocorrelation of the biodiversity index; this is particularly so for short distances.

Diversity index	r of test sets	CCC of test sets	N (p<0.05)
S	0.83 (≤ 0.03)	0.79 (≤ 0.04)	2
S_{ch1}	0.81 (≤ 0.03)	0.78 (≤ 0.04)	3
S_{ch2}	0.82 (≤ 0.03)	0.80 (≤ 0.03)	1
ES(20)	0.92 (≤ 0.02)	0.89 (≤ 0.02)	10
ES(25)	0.89 (≤ 0.03)	0.87 (≤ 0.03)	10
ES(50)	0.90 (≤ 0.02)	0.88 (≤ 0.03)	10
ES(100)	0.87 (≤ 0.03)	0.86 (≤ 0.03)	8
ES(150)	0.85 (≤ 0.02)	0.82 (≤ 0.03)	5
ES(200)	0.84 (≤ 0.02)	0.82 (≤ 0.03)	3
Δ	0.91 (≤ 0.04)	0.89 (≤ 0.04)	0
Δ^*	0.69 (≤ 0.05)	0.66 (≤ 0.05)	2
Δ^+	0.45 (≤ 0.09)	0.41 (≤ 0.09)	0
H'	0.91 (≤ 0.02)	0.89 (≤ 0.03)	10
H_{ch}	0.92 (≤ 0.02)	0.90 (≤ 0.02)	10

Table 3.3. Model performance of the ten selected models: 1) Average Pearson product-moment correlation coefficient of the test sets (±SE); 2) Average concordance correlation coefficient (CCC) of the test sets (±SE); 3) Number of models (N) with normally distributed residuals.

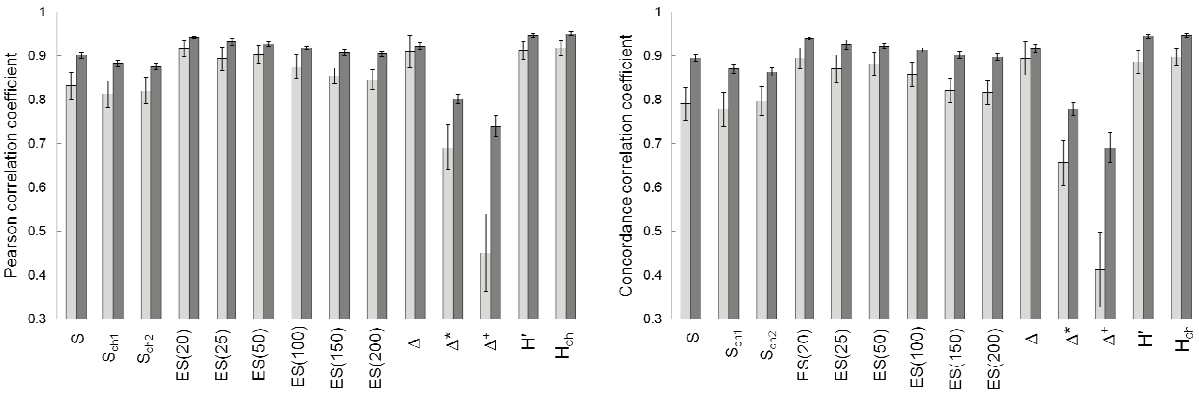

Fig. 3.2. Pearson product-moment correlation coefficient (left) and concordance correlation coefficient (CCC) (right) for observed and predicted values for the independent test set (light grey bars) and the learning set (dark grey bars) for the 10 selected models (±S.E.).

Input variables contribution

Two aspects of variable contribution were considered: the importance of a variable to the model output and the sense in which a variable contributes to the output. Since for each biodiversity index 10 models were selected, the contribution of the environmental variable was averaged over these models. We repeated the same analyses with only those models without significant SA in the residuals and found no differences for the variable contribution for all the diversity indices, except for taxonomic diversity. Thus, the networks of the latter are unstable; therefore, this index was excluded from the discussion.

For the first aspect, i.e. the importance of the variable, the three methods ranked the important predictor variables in the same way, thus indicating a well-calibrated network (Gevrey *et al.*, 2003). The Perturb method, however, differentiated the important variables more clearly, due to higher relative contributions. Some general patterns for the different diversity indices could be inferred: high relative contributions were found for sand and silt-clay (Fig. 3.3) except for Δ^+. Other important factors were the minimum total suspended matter, and sand extraction.

The second aspect, i.e. the sign of the influence of the environmental variable, was deduced by the Profile and the Modified Profile method. Analyses of the profiles did not show optimal intermediate values for the different environmental variables. In general, the biodiversity indices showed a positive correlation with the fraction of sand, the average current velocity at the bottom layer (Umea), the minimum bottom shear stress (Bstri), the maximum salinity, and intensity of sand extraction. Conversely, a negative effect could be discerned for clay, the minimum total suspended matter, and the average chlorophyll content.

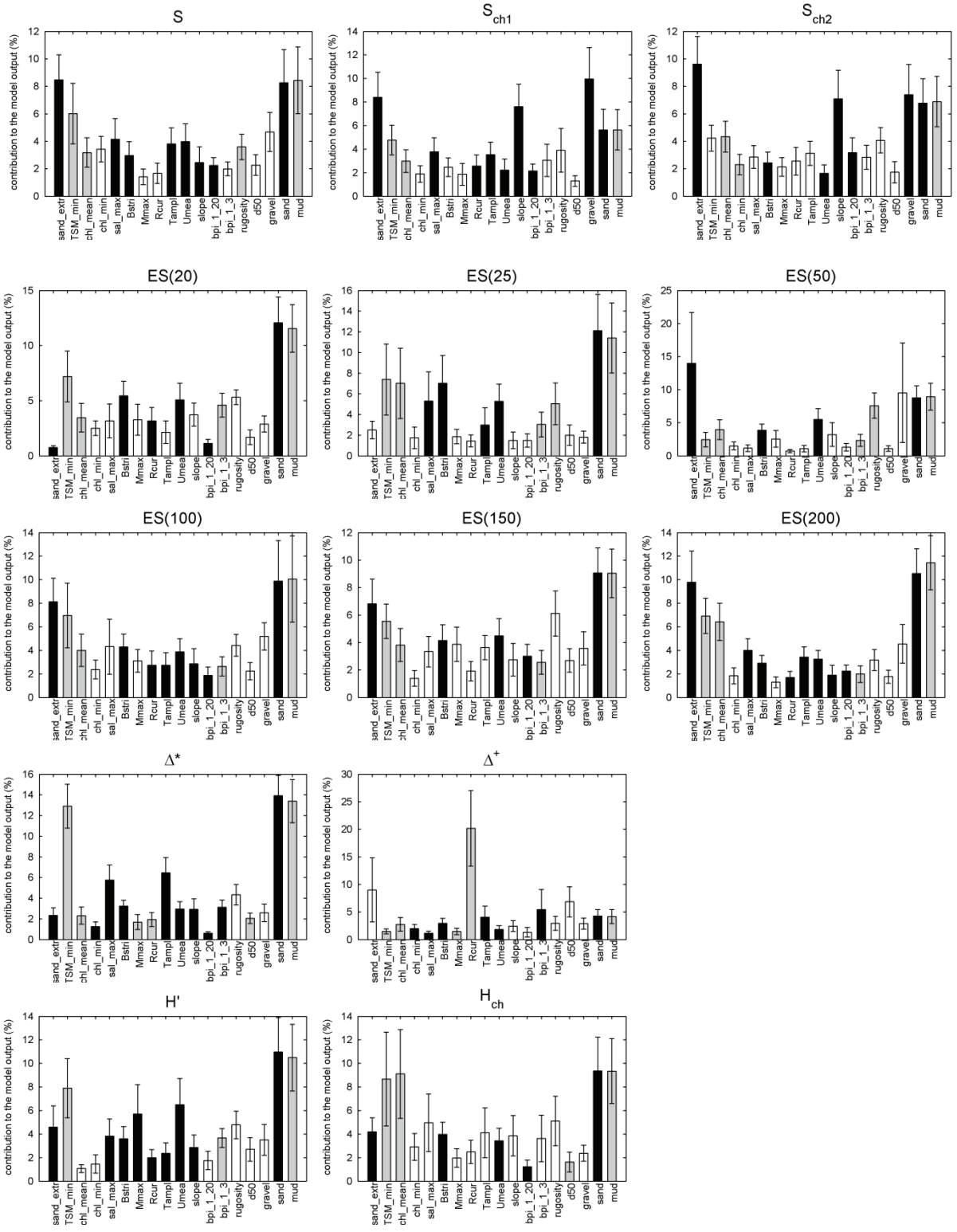

Fig. 3.3. Contribution of the environmental variables to the diversity indices according to the Perturb method averaged over the 10 selected models (±S.E.). The black bars represent those variables which are positively correlated with the diversity index for at least 75% of the models according to the Profile and the Modified Profile method. The grey bars imply a negative correlation, while the white bars indicate no straightforward positive or negative correlation with the biodiversity index.

DISCUSSION

Biodiversity indices and autocorrelation

According to Segurado *et al.* (2006) strongly autocorrelated environmental predictors, together with highly autocorrelated species distributions, lead to an inflation of the predictive power of the model. Our research supports this hypothesis, witnessing the strong correlation between Moran's I and the CCC of the test set. We minimised this problem by keeping adjacent samples in the same subset. However, as populations and environmental variables tend to be autocorrelated at all scales, the spacing out of samples will reduce, but never fully eliminate SA effects (Segurado *et al.*, 2006). Obviously, if a clear relationship between some environmental variables and diversity exists, the latter is easily predicted. However, even without such strong relationships, the inflation of the explanatory power for spatially autocorrelated variables makes them more likely to be selected in the final models (Segurado *et al.*, 2006). Consequently, causal relationships are difficult to distinguish and interpretation of ecological models has to be carried out with caution. Therefore, we believe that expert knowledge is essential for interpretation of the models.

The 14 diversity indices represent different aspects of diversity: species richness, evenness and taxonomic diversity, although these aspects are not equally represented in these indices. For example, the expected number of species is influenced by both evenness and species richness and with increasing n, the aspect of species richness will prevail; in general, sample processing in meiobenthic studies involves subsampling by randomly picking out 200 nematodes. Consequently, the value of ES(200) will approximate the number of species S found in the sample. Conversely, for low n-values, evenness contributes more to ES(n). Regarding the taxonomic indices: taxonomic diversity (Δ) is strongly correlated with the evenness parameter, Shannon's index of diversity (Clarke and Warwick, 2001), while taxonomic distinctness (Δ^*) represents more pure taxonomic relatedness and is therefore less associated with evenness. The average taxonomic distinctness (Δ^+) and species richness (S) are based on occurrence data and are thus independent of evenness. The extrapolated values of the species richness, S_{ch1} and S_{ch2}, are affected by rare species, found only once or twice in a sample, and are therefore only slightly influenced by evenness. With this in mind, it is clear from Table 3.2, that evenness, reflected in H', H_{ch}, ES(20), ES(25) and Δ, exhibits the strongest spatial dependence between the samples, followed by species richness, and pure taxonomic relatedness has the lowest SA.

Strong spatial autocorrelation of the biodiversity indices can be attributed to either the physical forcing of environmental variables or to community processes (Legendre, 1993). Since the model residuals show little or no spatial dependency, we believe that the environmental variables attributing to the spatial dependency of the diversity indices are well represented in the dataset. However, environmental factors may interact with biotic processes to generate these regional patterns. Disturbance of the seafloor by abiotic factors can affect diversity by regulating levels of competition, predation, and physiological stress

(Levin *et al.*, 2001). Since evenness has the strongest SA this aspect appears to be more strongly influenced by the physical forcing of the environment.

The residuals from the indices strongly influenced by evenness (H', H_{ch}, ES(20), ES(25) and ES(50)) are normally distributed, indicating that the models explain the variation in the dataset. With increasing n for ES(n), the number of models with normally distributed residuals decreases. Few models of species richness indices and taxonomic diversity meet this condition. Remarkably, not a single model of taxonomic diversity had normally distributed residuals, while this index is strongly influenced by evenness and showed a high SA and high accuracy (CCC = 0.89). Thus, part of the variation is still not explained by the model. Other modelling techniques could be more adequate in this case or other fine scaled abiotic or biotic variables may influence these diversity indices. These variables need to be fine scaled, since little or no SA remains in the residuals.

The two occurrence based indices, species richness S and the taxonomic distinctness index Δ^+, display a dissimilar pattern: S is strongly autocorrelated for nearby samples (Moran's I = 0.68), while Δ^+ shows much less resemblance for replicate samples (Moran's I = 0.26). Hence, two samples within a certain range will have similar evenness and about the same number of species, but taxonomically the communities in both samples show less resemblance, suggesting that the species inhabiting this area can be quite different.

Input variables contribution

In contrast with linear models, the contribution of each environmental variable to the output of the neural network is not easy to decipher. Although, it is possible to determine the overall influence of each predictor variable, interactions between the variables are difficult to interpret (Olden and Jackson, 2002) and different ranges of the abiotic variables may result in different outcomes. Moreover, causal relationships are difficult to unravel in biogeography and interpretation of ecological models has to be done cautiously. Validating these models should therefore include expert knowledge, and results from previous research can help in identifying relevant relationships.

The 'habitat heterogeneity hypothesis' is one of the keystones in ecology. It states that structurally complex habitats may provide more niches and diverse ways of exploiting the environmental resources and consequently increase species diversity. Habitats with a large degree of vertical and horizontal micro-environmental habitat heterogeneity seem to enhance diversity (Bazzaz, 1975; Tews *et al.*, 2004). Increasing sand and gravel content have a positive effect on species diversity of marine nematodes (Vanreusel, 1990; Steyaert *et al.*, 1999; Vanaverbeke *et al.*, 2002): higher diversity is associated with clean, coarser sand rather than with fine-grained coastal sediments. This positive effect of sand and gravel is possibly due to the larger interstitial space and consequently a higher number of microhabitats (Heip *et al.*, 1985; Vanaverbeke *et al.*, 2004b). This theory is confirmed by our models (Fig. 3.3): all diversity indices (except for Δ^+) are strongly influenced by the sediment composition: positively by the sand fraction and negatively by the silt-clay fraction. The

gravel content appears to increase the number of rare species found only once or twice in a sample, resulting in a strong contribution to S_{ch1} and S_{ch2}.

The adverse effect of organic input on nematode diversity has been reported before (Steyaert *et al.*, 1999), and is probably due to the anoxia resulting from eutrophication. Our analyses indicate that persistently high minimum TSM is more detrimental to the biodiversity of the nematode assemblages than accidently high inputs or average high inputs. The causal effect should be further investigated, but if true, it suggests that further efforts to reduce eutrophication could increase diversity of the meiobenthic community.

The hydrodynamic properties appear to be of less importance to the diversity, although generally they have a small but recurring positive effect on most of the diversity indices, except for the adverse effect of the residual currents (Rcur) on the taxonomic distinctness. The positive influence of an intermediate mechanical disturbance on the nematode diversity has been reported before (Vanreusel, 1990; Gage, 1996). However, the underlying reason is not clear and several hypotheses can be raised: (1) stronger currents allow species to disperse and colonise new patches; (2) new available patches are created which can be colonised by opportunistic species; (3) mild physical disturbance may switch systems where competitive exclusion would lead to reduced richness to systems where disturbance-mediated competitive coexistence occurs (Menge and Sutherland, 1976); (4) deposition of fine particles is prohibited by stronger currents, resulting in a higher sand fraction, more oxygen and habitat heterogeneity. In fact, a Pearson product moment correlation coefficient of -0.55 is found between silt-clay and the minimum bottom shear stress. However, other research associates high turbulence with low diversity (Heip *et al.*, 1985). Clearly, no cause effect conclusions can be drawn based on this circumstantial evidence.

Sand extraction, the only anthropogenic variable, shows a strong positive relation with most of the diversity indices. Again, the causal relationship is not clear and may be indirect: firstly, sand extraction occurs preferably at locations with coarse sand and is therefore a proxy for habitat heterogeneity and secondly, hydrodynamical forces are stronger on top of the sandbanks causing an increase in the biodiversity by intermediate mechanical disturbance.

The maximum salinity tends to have a positive influence on the diversity. According to literature however evenness (H') reaches a peak in the poly- to mesohaline zones, thus a salinity between 5 and 30 (Heip *et al.*, 1985). The minimum value in our dataset was 29.5; therefore, this hypothesis could not be tested. The low salinity values are found near the Scheldt estuary, a silty environment with high total suspended matter concentrations. Consequently, the influence of salinity is not straightforward and it may be a confounding factor.

CONCLUSIONS

Neural networks, often seen as a flexible but 'black box' tool, can be successfully implemented in ecological studies; different aspects of diversity are accurately predicted,

and those biologically relevant abiotic factors, such as sediment characteristics and organic input, are selected. Still, cautious interpretation is important because association does not necessarily imply causation and spatial autocorrelation may amplify or blur true ecological relations. Moreover, it is advisable to include several models in the final analysis, since a suboptimal model may be selected, resulting in faulty variable contributions.

Diversity maps are a useful instrument for decision makers in delineating high diverse areas. Based on our results and with geostatistics reliable maps of evenness and species richness of the nematode community of the BCS can be created. Together with diversity maps from other taxonomical groups, such as the macrobenthos, these maps could delineate areas of high biological interest.

There is a strong relationship between the predictability and the SA of a diversity index. The high SA of the diversity index can be attributed to the environmental variables in the model, since the residuals of the relevant models showed little or no SA anymore. It is clear that aspects of biodiversity, such as evenness and species richness show large scaled patterns in contrast to taxonomic distinctness. The high spatial variability of the latter makes it less suitable for area covering predictions. However, the explanation of this high variability remains an intriguing question. What factors determine taxonomic diversity? Different factors may attribute to this variability: (1) anthropogenic disturbance as suggested by Clarke and Warwick (2001); (2) small scale abiotic factors or (3) community processes such as competition and predation. Unravelling these factors is a challenging task but could shed light on fundamental questions such as the origin of the high meiofaunal diversity.

ACKNOWLEDGEMENTS

This research is funded by the BOF-fund of Ghent University (Nr. 01D02905) and 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: the Federal Institute of Economics (sand extraction data), the Management Unit of the North Sea Mathematical Models and the Scheldt estuary (MUMM, www.mumm.ac.be) for the current properties. ESA and MUMM/RBINS are acknowledged for providing and processing MERIS data (chlorophyll and TSM data, www.mumm.ac.be/BELCOLOUR), the Renard Centre of Marine Geology (RCMG, www.rcmg.ugent.be) of Ghent University for the oceanographic and sedimentological data. Special thanks to the Flanders Marine institute (VLIZ, www.vliz.be) for the help with building the biological database. This research was conducted within the MANUELA framework (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 publication is Contribution Number MPS-09018 of MarBEF. This research was also supported by the GENT-BOF Project 01GZ0705

Biodiversity and Biogeography of the Sea (BBSea). The authors would like to thank Emily Dolan and Jeroen Ingels who did an excellent job on revising the English grammar and style. Comments of two reviewers greatly improved the quality of the manuscript.