CHAPTER 8

GENERAL DISCUSSION

CHAPTER 8

GENERAL DISCUSSION

MODELLING

Modelling ecological and spatial data

Ecological data are often bulky, non-linear and complex, showing noise, redundancy, internal relationships and outliers (Park *et al.*, 2003). For this reason, in the last decade more complex datamining techniques have gained popularity for ecological modelling for different applications (Lek and Guégan, 1999).

Here, we modelled primarily the relationship between environmental variables and 1) the local diversity of the nematode community and 2) the probability of occurrence of specific species. Depending on the type of data and the goal of the model different modelling techniques can be applied. In Table 8.1 the main features of linear models, ANNs and Maxent are compared. The main difference between conventional statistical modelling techniques and more flexible data mining techniques, such as ANNs, lies in the basic assumption of linear models: linear models assume a linear or link linear relation between the independent and dependent variables. If the relationship is expected to be more complex or unknown ANNs are preferred. Our goal was to compare the predictability of the different aspects of diversity and thus create predictive models which explain as much the variation in the data as possible regardless of the complexity of this relation. Since a lot of samples were available for the prediction of the local diversity (209 samples) a powerful tool such as ANN could be applied. However, for habitat suitability modelling ANNs would have been a less adequate choice since the number of data points was generally much lower (between 5 and 106 samples for the nematode species and 30 samples on average). In that case, a generative approach may be more appropriate. These models can yield good predictions even with small sample sizes (Ng and Jordan, 2001). Generally, absence of a species is difficult to establish, in such cases a presence-only approach is more appropriate.

All the input data used in this study is spatial data which may generate problems when being analysed: 1) sampling of specific regions or specific environments may result in a sampling bias, which may result in accepting species models for which the good predictive capacity might be entirely attributed to this sampling bias. 2) spatial autocorrelation may contribute to an inflated predictive power of the models. These issues are often overlooked or ignored in other studies (Dormann, 2007). Therefore, the current modelling protocols were refined to address these aspects.

In this thesis, we focused mainly on three issues: spatial autocorrelation, preferential sampling and overfitting. In the previous chapters these issues were tackled in different ways, in this concluding chapter we will give a short review on these different issues and the techniques we used.

	Linear models	ANN	Maxent
Purpose	Model linear or link linear relations between the independent and dependent variable	Model any relation between the independent and dependent variable. ANNs can approximate any mathematical function.	Model the relation between the environmental variables and the probability of occurrence of a species
Data needed	Number depends on the strength of the relation and complexity of the model	Large number of data. Guideline: the number of data points should be at least 10 times the number of weights in the network (Fernandes and Lona, 2005)	Uses presence-only data and can work with as few as 5 occurrence localities (Pearson *et al.*, 2007)
Approach	Discriminative approach	Discriminative approach	Generative approach
Ease of use	+	-	+
Interpretation of the model	Relation between independent and dependent variables is clear	Relation is not clear (Black box). Relation can be revealed in different ways (Gevrey *et al.*, 2003)	Relation is revealed by λ-values (and response curves).
Overfitting	Less prone to overfitting, but may not capture all the variation which can be explained by the independent variables.	Prone to overfitting. Can be reduced by cross-validation & early stopping.	Prone to overfitting. Can be reduced by cross-validation and ℓ_1-regularisation
Preferential sampling (Sampling bias)	Sampling bias has an effect on the output	Sampling bias has an effect on the output	Sampling bias has a strong effect on the model output (Phillips *et al.*, 2009)
Data output	One solution	Depending on the initial weights different solutions can be found	One solution

Table 8.1 Overview of different aspects of linear models, ANNs and Maxent.

Techniques

Neural networks

Since relations between biodiversity and the environmental variables may be complex, flexible learning techniques like ANNs are adequate to study these relationships, since they can simulate any continuous mathematical function and can therefore describe complex ecological functions (Olden *et al.*, 2008). Two potential drawbacks of this methodology are that they are susceptible to overfitting and they act as a 'black box' (Lek *et al.*, 1996a). In this thesis, we dealt with both issues: overfitting is tackled by applying a 10-fold cross-validation and early stopping (see further) and the contribution of each environmental variable to the output is revealed by applying three methods: two known in literature: the Perturb method (Yao *et al.*, 1998; Gevrey *et al.*, 2003) which gives information of the contribution of the variable to the model, and the Profile method (Lek *et al.*, 1995, 1996a,b; Gevrey *et al.*, 2003) which provides information on both the importance and the sign of the environmental variables. A third technique was applied to check the validity of the previous two techniques (Chapter 3). This Modified Profile method combines aspects of the Perturb and the Profile method. The three methods revealed the contribution of the environmental variables to the models. These results are in accordance with previous knowledge on the diversity of nematode communities and allow for a generalisation of this knowledge on a broad geographical scale. Thus, notwithstanding the complexity of the models, ANNs are able to select the relevant environmental variables contributing to nematode diversity.

Maxent

Numerous methods and software packages have been developed to model a species fundamental niche (Guisan and Zimmermann, 2000; Guisan and Thuiller, 2005; Elith *et al.*, 2006; Guisan *et al.*, 2007). These models are based either on presence/absence data or on presence-only data. For the data at hand the choice of a presence-only modelling technique is valid for several reasons: all the modelled species have a patchy distribution. Hence, absence does not necessarily mean that the habitat is unsuitable for the species. Moreover, nematode species are inconspicuous and the presence of a species may be easily overlooked. In the case of invasive species, such as *Ensis directus*, the species may not have reached its equilibrium distribution yet in the area under investigation. Therefore, Maxent, a presence-only niche modelling technique is chosen. Moreover, Maxent ranks within the best performing modelling techniques (Elith *et al.*, 2006; Guisan *et al.*, 2007; Ortega-Huerta and Peterson, 2008; Benito *et al.*, 2009).

Generalist species, which can survive in a broad range of environmental conditions are generally hard to model (Segurado and Araújo, 2004; Evangelista *et al.*, 2008). Recent research emphasises the importance of common species to ecosystems (Gaston and Fuller, 2008) and identifying regions where a species can occur in high densities can be useful for nature conservation (e.g. *Lanice conchilega*) or for fisheries (e.g. *Ensis directus*) purposes. Since HSMs are using observational data (presences), the modelling algorithm cannot differentiate between high and low densities of the species. By applying thresholds on the relative abundance of nematode species and the total density of *L. conchilega*, we were able to successfully delineate regions where these species can thrive and occur in high relative abundances or high numbers, and environments where the species occurs in low relative abundances or numbers (Chapter 6).

Null models

A null model is a model which generates a simulated pattern under restricted ecological or evolutionary assumptions (Gotelli and Graves, 1996). These null models can be a powerful tool for testing non-standard hypotheses about patterns in ecological data, while using observed patterns. We used null models in two ways: for testing whether the observed community structure is significantly different from random (Chapter 2) and for revealing the presence of preferential sampling, overfitting and spatial autocorrelation (Chapter 5 and 7).

The advantage of null models is that they provide flexibility and specificity which cannot be obtained with conventional statistical analyses (Gotelli, 2001). Moreover, these models bear close resemblance with the natural conditions by keeping different aspects of the data constant which allows mimicking natural conditions.

In this thesis, these null models have proven their usefulness in resolving complex ecological issues such as revealing preferential sampling in the environmental space and discovering aggregation patterns in replicate samples. It should be noted that the required calculation time to build several hundreds or thousands of null models is a limiting factor.

Geostatistics

Geostatistics is an interpolation technique for spatial analyses taking into account spatial autocorrelation. It can be applied to any dataset showing a spatial structure and which is not purely random. Geostatistics require a lot of data to allow reliable mapping: the spatial dataset should consist of at least 100, and preferably 144 data points (Webster and Oliver, 2007). In our case, we had information about 153 different stations. However, including the replicate samples increased this number to 557 replicates. Replicate samples have the same geographical coordinates, but by randomly adding small variations within the range of a few meters, the replicate samples allowed a good estimation of the local variability of the measured variable (nugget). Excluding these replicate samples from the analysis revealed no spatial pattern in the residuals of the model. In our case, regression kriging clearly outperforms ordinary kriging, since a large part of the spatial variation can be explained by the environmental variables. However, a higher number of observations would be advisable, since the resulting maps shows local patches due to the large distance between some observations.

Modelling issues

Spatial autocorrelation

The issue with spatial autocorrelation (SA) is that it violates the original assumption of independent sampling in modelling. This may result in different adverse effects: it may be an important source of bias (Segurado *et al.*, 2006) and inflate type I errors (Dormann *et al.*, 2007). SA was shown to represent a serious problem for niche-based species' distribution models. Significance values were found to be inflated up to 90-fold (Segurado *et al.*, 2006). Moreover, the ignorance of SA may lead to the inversion of the observed pattern of an environmental variable (Kühn, 2007) and even cross-validation may not be able to reveal the inflating effect of spatial autocorrelation if no measures are taken (Telford and Birks, 2005). Not including SA may result in the choice of an inappropriate model with irrelevant environmental variables and overly optimistic estimates of the predictive power of the model. This leads to the use of models with no predictive power but with good performance statistics (Telford and Birks, 2005). Several methods exist to evaluate or prevent the potential influence of spatial autocorrelation. We applied four techniques: checking the spatial autocorrelation in the model residuals, applying spatially explicit models, applying geostatistics and spatially separating test and training sets during cross-validation.

The first methodology was applied in Chapter 3: the presence of spatial autocorrelation in the residuals of the artificial neural network models is evaluated by Moran's I (Moran, 1950). The advantage of this method is that it can be applied to any modelling technique. However, when spatial autocorrelation remains in the residuals, the validity of the model is questionable and other spatially explicit techniques should be applied.

Many spatially explicit techniques have been developed (Dormann *et al.*, 2007), such as spatial autoregressive models and generalised least squares. The latter technique was applied in Chapter 4 where we compared the performance of generalised least squares (GLS) (a spatially explicit modelling technique) with ordinary least squares (OLS) (a non-spatially explicit model). The GLS models slightly outperform the OLS models. However, the GLS models were not able to explain all the spatial variation in the data, since the variogram still shows some spatial pattern in the residuals (Fig. 4.3). Thus, even with spatially explicit models an amount of spatial variation may remain in the residuals of the model.

If mapping is desired, geostatistics uses the information about the spatial structure of the data (range, sill and nugget) to improve the prediction of data points within the range of the sampling points (Chapter 4). The latter methodology is rather labour intensive and at least 100 data points are needed to allow a reliable estimation of the variogram (Webster and Oliver, 2007), which is a high number in meiofaunal research.

Another way to reduce the influence of SA on the generalisation performance of the model, is to spatially separate test and training set during cross-validation. The estimation of the predictive power of a modelling technique assumes that the test sites are independent of the training sites. Cross-validation in the presence of spatial autocorrelation seriously violates this assumption (Telford and Birks, 2005). When the test and training set are spatially separated, the modelling algorithm will be trained in the direction of a general relationship (Telford and Birks, 2005). In Chapter 5 and 7 the training and test sets were spatially separated. This method may be more appropriate when a modelling technique, which does not explicitly incorporate spatial autocorrelation, is applied to a spatially biased dataset. In theory, the test and training set should be out of the range of the spatial autocorrelation. However, this is often practically unachievable. For instance, when predicting the diversity indices, both datasets should be at least 42 km apart. When applying a 5-fold cross-validation this would imply that the data points of the five subsets should be 42 km apart, which is practically unachievable for the region under study. For instance, for one third of all nematode species it was impossible to separate the data in five subsets which are at least five kilometre apart (Chapter 5).

Concluding, depending on the dataset and modelling technique used, one of these techniques can be applied when handling spatial data:

- Investigation of the **spatial autocorrelation in the model residuals** can only be used post modelling to check the validity of the model, but does not provide solutions to problems association with autocorrelation;
- A **spatial separation of test and training set** can be applied when a predefined method is used, as was the case with Maxent.
- Methods which directly incorporate spatial autocorrelation, may be preferred when **spatially explicit models** are developed. These techniques are the most straightforward way to handle SA.
- When mapping is required and enough data points are available, **geostatistics** are a useful tool to increase the accuracy of the maps.

Preferential sampling

Preferential sampling or sampling bias occurs when some areas are visited more than others. This may be due to the higher accessibility of the area or a biased sampling design. In this way, some parts of the environment may be undersampled or not sampled at all. In the latter case a correct prediction for the unsampled environment is impossible. If only a limited number of samples are available from a certain environment, a down weighting of the oversampled region can be suggested. When the degree of underrepresentation is small, the sample can be treated as a reasonable approximation of a random sample (Glasser, 2008).

In this study we investigated the presence of preferential sampling in two ways: by calculating the declustered mean for a given cell size (Chapter 4) and by applying null models (Chapter 5 and 7).

The declustered mean is calculated by dividing the geographical area into cells with the same size and for each cell the mean of the measured variable is calculated. The global mean is then calculated from these cell means. This is repeated for different cell sizes. If the global mean shows a minimum or a maximum for a certain cell size, the data should be declustered. This can be done by replacing the measurements in one cell by the average value in that cell or by assigning weights to the measurements. If no distinct minimum or maximum is observed, the observations are not preferentially clustered. This methodology was applied during the pre-modelling analysis in Chapter 4 (The analysis was not discussed in this chapter since there was no need for declustering): spots with higher or lower values of the diversity index were thus not preferentially sampled.

Another way to reveal the presence of preferential sampling is by the use of null models (Chapter 5 and 7): data points are randomly selected from the complete area and randomly selected from the sampled stations. Based on these subsets two models are developed and this is repeated several times (e.g. 500 times). If the predictive performance of the model based on the random samples from the sampling database is significantly better than the performance of the model based on the samples of the total area, preferential sampling is present. This method does not only detect the presence of preferential sampling, but it also indicates which species models can be considered being significantly different from random.

Although both techniques may reveal the presence of preferential sampling, they reveal different aspects of preferential sampling. The difference between these types of preferential sampling is often ignored. In fact three types of preferential sampling can be distinguished: preferential sampling of the dependent variable (Van Meirvenne, 2007), preferential sampling in the geographical space and preferential sampling in the environmental space (Pearson, 2007). Preferential sampling of the **dependent variable** can be revealed by calculating the declustered mean, as areas with low or high values of the measured variable are more intensively sampled. Preferential sampling in the **environmental space** can be detected by the use of null models. The null model based on random data from the sampling database will perform better than the null model based on random data from the total area if the sampling database holds an environmental bias. Preferential sampling in the **geographical space** does not necessarily inflict modelling issues since preferential sampling in the geographical space may still result in good sampling in the environmental space, which is actually used to build the model (Pearson, 2007).

Preferential sampling of the environmental space or of the dependent variable may also result from a different research question: for instance if a scientist is interested in identifying regions with high biodiversity, he/she may sample species rich areas more intensively although the environmental variables in these areas might fluctuate considerably. Thus, this will result in a bias of the measured variable, but not in an environmental bias. On the other hand, if a researcher is for instance interested in one particular species, and it is known in which environment this species is generally found, he/she might preferentially sample this environment, although the measured value of the dependent variable may strongly fluctuate e.g. due to a patchy distribution. The latter was clearly the case for the 2010 sampling campaign of *Ensis directus* (Chapter 7). In this particular case, the model showed an excellent performance (AUC=0.93) but the species model did not perform better than a random model. Thus, it is crucial to investigate the presence of preferential sampling in the environmental space, since it may not only exaggerate the performance of the model, but it could lead to restricted applicability of the model as the model can only fit to that portion of the environment which is included in the observations. Consequently, it can only identify a part of the actual and potential distribution of the species. Moreover, preferential sampling has a stronger effect on presence-only models than on presence/absence models (Phillips *et al.*, 2009). Thus, extra care is essential when working with Maxent.

Pitfalls concerning modelling: Overfitting

Overfitting occurs when a model is overly complex and the model fits the data points too closely. In that case, random error is modelled, rather than a meaningful underlying relationship. These models generally have poor predictive abilities and only describe the data at hand. To avoid overfitting, several techniques have been developed which either penalise overly complex models, or test the generalisation ability of the model by testing its performance on unseen data. In our research four techniques were applied: k-fold cross-validation, the use of a single validation test, ℓ_{1}-regularisation and early stopping.

When applying k-fold cross-validation, the data is split in k datasets. Each set is once used as a test set, while the rest of the data is used as training data. The global performance of the model is calculated by averaging the performance factor of the test set of the k models. Thus, the true error is estimated by calculating the average error rate. In this way, different models with different features and parameters can be compared easily. The advantage of a k-fold cross-validation is that all the samples are used for both training and testing. The choice of k depends on the available calculation time and the available data. Generally a 10-fold cross-validation gives good results concerning the bias and variance on the accuracy estimation of the models (Kohavi, 1995). For Chapter 3 a 10-fold cross-validation could be applied since a lot of data were available (209 samples) and the fast Levenberg-Marquardt training algorithm is used (Beale *et al.*, 2010). However, for the habitat suitability modelling (Chapter 4 to 7) the number of data points varied between 5 and 106 and the modelling speed seriously dropped with increasing number of samples. Therefore, a lower number of folds were applied.

Alternatively, if enough data is available and the calculation time of a single model is labour intensive and time consuming, validation of the different models can be done with a single test set which is used at the completion of the modelling procedure. This procedure was followed in Chapter 4.

Early stopping can only be applied in case the modelling technique includes an iterative optimisation procedure and when enough data is available. Here, the data is split in three subsets: a training set, a validation set used for early stopping and an independent test set to compare the performance of the final models. During each training cycle in the modelling algorithm, the error of the validation set is compared with the error of the validation set during the previous cycle. If the error on the validation set starts to increase, the model is starting to overfit and the training cycle is interrupted. The early stopping technique was used during model fitting of the ANNs in Chapter 3 in combination with a 10-fold cross-validation.

A fourth method to reduce overfitting is by using regularisation: a penalty term is added which penalises complex models having many parameters. The aim of regularisation is to trade off model fit and model complexity (Elith *et al.*, 2011). The Maxent software includes a ℓ_1-regularisation parameter which is closely related to the Akaike's Information Criterion (AIC, Akaike, 1974) another penalty criterion for complexity (Elith *et al.*, 2011). Our results from Chapter 5 indicate that with the default value of the ℓ_1-regularisation parameter the model still tends to overfit. In that case parameter tuning to further reduce overfitting is needed (Phillips and Dudík, 2008). This can be achieved by either adjusting the l_1-parameter (Holt *et al.*, 2009), or by parameter and feature selection by the use of cross-validation (Chapter 5, 6 and 7). We reviewed 53 papers where Maxent was used, although the majority mentioned the use of a test set and cross-validation, some papers did not. In the latter case the estimate of the AUC may be unrealistically high and moreover, the presence of overfitting cannot be detected.

The choice of the technique to reduce overfitting depends on the modelling technique and the data at hand. Early stopping is commonly used in machine learning and with neural networks, while regularisation is integrated in the Maxent software. However, our results show that an additional cross-validation is further needed to fine-tune the model. Thus, relying on a single technique is not advisable and cross-validation proved to be a valuable way to create a good generalising model. If enough data is available, a single test set can be used, while cross-validation has the advantage that all the data is used during model development.

Overfitting and spatial autocorrelation

Interestingly, overfitting may also result from spatial autocorrelation (Telford and Birks, 2005): if the samples of the training and the test set are autocorrelated, an overfitted model may still have good predictive power and the predictive performance of the model may be exaggerated. Spatially separating test and training set is one step in reducing the consequences of both issues. This effect is possibly at work for the *Ensis directus* modelling: when the datasets are not spatially separated a complex model is selected by the cross-validation, while a more parsimonious and straightforward model is selected when the datasets are spatially separated (Chapter 7). The final habitat suitability models for the nematode species confirm this (Addendum 3): on average less environmental variables are selected in the final model when the distance between test and training sets is increased. Thus, when working with spatial data a combined approach of cross-validation and a spatial separation of the individual sets is advisable. Moreover, when a region is oversampled compared to other regions in the area (e.g. western region near the Belgian coast in the

MacroDat database), the algorithm designed to spatially separate the subsets (Addendum 5) will reduce the number of samples in the oversampled region, since the algorithm assigns the same number of samples (± 1) to each subset. In this way, the potential influence of preferential sampling on the model outcome is counteracted too.

Limitations to the models

Creating a model with a well-designed software package is often very easy. One only needs to enter the data in the correct format in a software package and within seconds a model is produced. However, a model can only be as good as the data it is based on (Pearson, 2007). If the data does not provide useful information, or biased information, then the model cannot provide useful information.

Species data

It should be mentioned that the data used in this study was compiled from different datasets from different researchers. Methodological differences between researchers may increase the heterogeneity in the data. There may be individual differences between researchers such as differences in sampling techniques, subsampling effort, identification level of the species and other factors. Interpreting the analyses resulting from this data may not be without risk (Soetaert and Heip, 1995). Therefore, the data should be standardised or the modelling technique should be developed in such a way that the personal influence of the researcher is eliminated. Here, we did this by restricting the swapping algorithms of the null models to replicate samples (Chapter 2) and by using standardised biodiversity indices (Chapter 3 and 4). Replicate samples are sampled and analyzed by the same researcher, for the same research topic and with the same sampling gear.

Environmental data

One of the outcomes of this research are the maps of the biodiversity indices (Chapter 4) and the habitat suitability maps of the nematode (Chapter 6, Addendum 3) and macrobenthos species (Chapter 7). These maps summarise the information captured by the model in an easy to use format. However, these models should be considered within the constraints of the models. The majority of the models were built with nine environmental data layers with a resolution of about 250 m. As such, it is very unlikely that all factors contributing to the species' niche or to the nematode diversity are incorporated in the model. In addition, data on small scale variations in the environmental data are not available to refine the models. Moreover, the environmental data extracted from the maps are not flawless. Little is known about the accuracy or local variance of these environmental maps. The environmental data are a snapshot of the actual situation. Many factors may alter the actual situation e.g. the position of the silt-clay deposits has changed the last century due to human activities (Fettweis *et al.*, 2009) . At the beginning of the 20th century layers of fresh silt-clay were found at the near shore area between Ostend and Zeebrugge, while nowadays they are concentrated in the area in front of Zeebrugge (Fettweis *et al.*, 2009). This is not a very strong shift, but it does indicate that species occurrences may shift over time. For the chlorophyll *a* data no clear trends has been found in the time frame 1975-1991 in the Southern Bight of the North Sea (De Cauwer et al., 2004). However, this does not exclude the possibility of changes in the future. The differences in species distribution due to these changes can be estimated by applying these altered environmental maps to the existing model.

Nematode communities change seasonally (Vincx, 1989b; Vanaverbeke *et al.*, 2004a; Franco *et al.*, 2008). The current HSMs predicts where a suitable habitat is found for the species based on annual data. If the species appears at a location at a certain moment in time, but is absent during other periods, the habitat will be assigned as being suitable for the species. Splitting up the data seasonally for both environmental data (chl a and TSM) and the species data would be an option. However, this would result in a strong data reduction for the species data and increased uncertainty of the environmental maps.

Missing data

Concerning the diversity maps (Chapter 4), an area covering map will inevitably incorporate regions where the model extrapolates for unknown environmental conditions in the original dataset. Consequently, these models should be treated with caution, especially in regions which are less visited, such as the Northern part of the study area.

Regarding the HSMs (Chapter 6 and 7), the data is always incomplete: in reality, species are unlikely to occur in all suitable areas. Moreover, the occurrence data will not reflect the complete range of environmental conditions suitable to the species. Therefore, these models should not be expected to predict the full extent of the actual or the potential distribution of the species (Pearson, 2007). Moreover, areas within the predicted suitable habitat may not be occupied by the species because of a patchy distribution (e.g. nematode species, *Lanice conchilega*), or because the species did not yet occupy the full extent of its potential distribution (e.g. *Ensis directus*) or because it is excluded from the area due to biotic interactions or environmental conditions not incorporated in the model.

Nevertheless, our maps offer valuable information which can be used for different purposes: conservation management, identifying diversity hotspots (Graham *et al.*, 2004) (Chapter 4 and 7) and for fisheries (Chapter 7). Other potential applications of HSMs exist in identifying the potential area of an invasive species, based on its original habitat (Thuiller *et al.*, 2005); modelling the impact of climate change on a species' distribution (Berry *et al.*, 2002); and identifying regions where the species is potentially present, but not yet observed due to insufficient sampling (Pearson *et al.*, 2007).

BIODIVERSITY

There is growing need on insights explaining patterns of biodiversity and knowledge on the local biodiversity of regions for nature conservation. In the introduction, we gave an overview on several hypotheses explaining biodiversity (Table 1.3). Some of these theories are based on large-scale processes, such as differences in latitudinal diversity and the importance of climatic stability, while other processes act on a small scale such as competition and biological disturbances. Other theories like the habitat heterogeneity hypothesis, can be applied to both small and large scale. It is generally accepted that biological processes are nested within physical processes (Levin *et al.*, 2001), but this does not necessarily mean that biological processes are always small-scaled (e.g. migrations) and physical processes broad-scaled (e.g. micro-topography) (Legendre and Legendre, 1998). Some of these hypotheses like island biography, stability-time hypothesis, climatic stability, historical explanations and latitudinal diversity hypotheses, cannot be tested with the data at hand. These hypotheses will therefore not be touched upon in this chapter. We focused on biological interactions, environmental factors including productivity measures (chl *a* and TSM), disturbance measures (current characteristics) and habitat heterogeneity (sediment characteristics and topological measures) on a limited time (1975-2010) and spatial scale (Southern Bight of the North Sea).

It should be stressed that no cause-effect relationships can be drawn from the data. Moreover, the environmental variables selected in the models may be a proxy for other variables which may have a more straightforward impact on the diversity. The actual testing of these theories should be done by experimental setups excluding other interfering factors. However, our findings are compared with results from previous research to check whether model outcomes can be used to corroborate theoretical ecological frameworks.

Biological interactions

The search for species assembly rules focuses on the influence of interspecific interactions. It has been claimed that competition is a driving factor for species to evolve (Dobzhansky, 1950; Dayton, 1971; Grassle and Sanders, 1973; Diamond, 1975). However, it is impossible to draw this conclusion from data collected in the field. First of all, establishing competition from a dataset is precarious. The data may hold structures resembling ‘assembly rules’ such as segregation and aggregation of species, but many other factors may be at work. Checkerboard patterns may be a result of species showing affinities for certain habitats (Gotelli and McGabe, 2002), or of historical evolutionary events. Moreover, even if competition has led to behavioural, distributional or morphological differences between species, this is hard to demonstrate in a present-day dataset, where in fact competitive interaction between species may be strongly reduced or even disappeared. Thus the current data may represent ‘the ghost of competition past’ (Connell, 1980). Moreover, demonstrating co-evolutionary divergence involves revealing resource partitioning and evolutionary character displacement, which in fact reduces present-day competition. There is abundant observational evidence for ecological character displacement in general (Pritchard and Schlüter, 2001) and even for nematodes this has been suggested (Wieser, 1960). Notwithstanding these examples, the evidence is incomplete and these examples should be further supported by evidence demonstrating that resource competition is actually present and is the mechanism driving divergence (Connell, 1980; Pritchard and Schlüter, 2001). This can only be shown in a carefully monitored experimental set-up. Disentangling cause and consequence is thus not possible with the short term data at hand and the null models should only be viewed as statistical tools to recognise non-random species distribution patterns (Gotelli, 2001). Although most ecological studies indicate the presence of less co-occurrence than expected by chance (segregated communities) (Gotelli and McGabe, 2002), our null models based on replicate samples (Chapter 2) point in the direction of aggregated patterns of nematode communities. This patchy and aggregated distribution of meiofaunal species is not new to science and it has been attributed to many different causes: microtopography (Hogue and Miller, 1981; Sun *et al.*, 1993; Blome *et al.*, 1999), the presence of biogenic structures and macrofauna (Reise, 1981; Braeckman *et al.*, 2011), food source patchiness (Lee *et al.*, 1977; Blanchard, 1990), and even (social) species interactions have been suggested for meiofaunal communities (Heip, 1975; Findlay, 1981; Chandler and Fleeger, 1987). This study thus reaffirms the presence of patchy and aggregated communities and shows that these are found on a broad spatial scale. However, this analysis does not exclude the presence of competitive processes on a smaller scale; segregated patterns have been found within samples based on depth slices (Joint *et al.*, 1982; Steyaert *et al.*, 2003). However, the cause of these segregated patterns can be attributed to either environmental changes or interactions between species.

Competition is mostly expected between species of the same trophic group (Fox and Brown, 1993). Therefore, we subdivided our data in the four trophic groups described by Wieser (1953). The null models revealed the same aggregated patterns for all the feeding types. Thus even within feeding types, we did not find segregated patterns. Again, this could be due to the relatively large size of a core in respect to the nematodes, small habitat differences between replicate cores or the coarse subdivision in the four feeding types of the original classification of Wieser (1953). More specific feeding types have been described (Moens and Vincx, 1997; Moens *et al.*, 2004). Unfortunately, this classification is not known for a lot of nematode species. Moreover, nematodes can display complex feeding behaviour (Postma-Blaauw *et al.*, 2005; dos Santos *et al.*, 2009) sometimes even related to the food availability (Giere, 2009), which complicates the subdivision in different feeding types.

Competition could also differ according to the environment: according to Schratzberger and Warwick (1998) lower competition for resources is expected in sands since it has generally higher disturbance levels. In contrast, biological and competitive interactions are more likely to occur in sheltered, more stable, muddy sediments with infrequent disturbances (Schratzberger and Warwick, 1998). On the other hand, Armenteros et al (2010) found no food limitation for nematodes in a natural muddy environment. Franco *et al.* (2008) showed that in sandy sediments chl *a* levels are much lower than in muddy environments and both in sandy and muddy sediments nematode biomass and densities increase after deposition of a phytoplankton bloom (Franco *et al.*, 2010) suggesting a food limited nematode community in other periods. Thus, whether food limitation, and thus more competition, can be expected in one of these contrasting environments is still unknown. This hypothesis was not tested since sediment data was scarcely present for the repeated samples in the database.

In conclusion, we did not find evidence of species interactions leading to less co-occurrence than expected by chance. The nematode communities reveal strong aggregated patterns, but the cause of these patterns cannot be revealed with the data at hand. Thus, our analysis does not refute the hypothesis that species interactions may structure nematode communities, but it does not support it either.

Habitat heterogeneity hypothesis

The 'habitat heterogeneity hypothesis' states that structurally complex habitats may result in a higher number of niches and may thus provide more ways to exploit the resources and consequently increase species diversity (MacArthur and MacArthur, 1961; MacArthur and Wilson, 1967). A large degree of vertical and horizontal micro-environmental habitat heterogeneity enhances diversity (Bazzaz, 1975). The positive influence of a heterogeneous habitat on diversity has been widely reported in terrestrial (Tews *et al.*, 2004) and marine environments (Levin *et al.*, 1986). More specifically, for the meiobenthic community, this hypothesis has been related to large scale habitat heterogeneity (Vanreusel *et al.*, 2010) and small scale habitat heterogeneity (Gingold *et al.*, 2010b). Increasing sand and gravel content are strongly related to a higher nematode diversity (Heip *et al.*, 1985; Vincx *et al.*, 1990; Vanaverbeke *et al.*, 2002; Vanaverbeke *et al.*, 2004b; Vanreusel *et al.*, 2010). Clean well sorted fine to coarse sands may contribute to habitat heterogeneity (Vincx *et al.*, 1990). These sediments harbour more microhabitats and sediment particles larger than 300 µm show more flat surfaces than smaller particles allowing a wider variety of bacterial colonies to colonise these areas (Giere, 2009). Moreover, it has been stated that intermediate grain sizes provide optimal space for most nematodes to move (*in* Abebe *et al.*, 2006). However, other confounding factors such as lower food availability, more disturbance and the absence of oxygen stress may also be associated with a higher sand fraction (Schratzberger and Warwick, 1998; Steyaert *et al.*, 1999; Franco *et al.*, 2008; Vanaverbeke *et al.*, 2011). The positive influence of small-scale habitat heterogeneity on species diversity has also been reported for the deep-sea (Tietjen, 1984; Tietjen, 1989).

Our models confirm the strong positive relationship between α-diversity and the sand and gravel fraction. Only the average taxonomic distinctness (Δ^+), which can be seen as the average taxonomic path length between any two randomly chosen species, does not exhibit this strong relationship (Chapter 3). However, even with this well established relation between the diversity and the sand fraction, these results do not provide evidence about the real cause of the high diversity in sandy sediments.

Influence of disturbance and productivity

Disturbance

The intermediate disturbance hypothesis (IDH) (Connell, 1978) states that the diversity will be maximised at an intermediate level of disturbance due to the elimination of strong competitive species, which allows co-existence of less competitive, more opportunistic species (Connell, 1978). The intermediate disturbance hypothesis has been supported for meiofaunal communities in freshwater environments (Witthöft-Mühlmann *et al.*, 2007) and on sandy beaches (Armonies and Reise, 2000; Gheskiere *et al.* 2004; Gingold *et al.*, 2010b). Intermediate biotic disturbances in sublittoral regions do increase the diversity (Austen *et al.*, 1998; Widdicombe and Austen, 1998). Physical disturbances revealed that nematode communities in muddy sediments follow the IDH hypothesis and have the highest diversity at intermediate disturbance levels, while in sandy sediments these communities show more resilience and recover more quickly, probably because these species are adapted to more disturbed natural environments (Schratzberger and Warwick, 1998). However, other benthic research does not support the IDH (Van Colen *et al.*, 2010) and it has been shown that only 20% of the research done on the IDH actually supports the hypothesis (Mackey and Currie, 2001). Another drawback of the hypothesis is that it is a conceptual model and the intensity of disturbance and the nature of disturbance are not clearly defined (Svensson *et al.*, 2010). The effect of disturbance on species richness may depend on the specific combination of frequency and area of the disturbance (Svensson *et al.*, 2010). Moreover, natural and anthropogenic disturbance may result in different effects on the exposed populations (Schratzberger *et al.*, 2009). In the deep-sea, differences in current velocities do not seem to have an effect on nematode diversity (Lambshead *et al.*, 1994).

With our data it is hard to test the IDH. However we do notice that biodiversity shows a positive correlation with the following disturbance variables: the average current velocity at the bottom layer, the minimum bottom shear stress and the intensity of sand extraction. However, the analysis did not point out an intermediate optimum. Moreover, these environmental variables may be related to the sand and silt-clay fraction: lower values of the current velocity and of the minimum bottom shear stress allow deposition of mud particles and are thus related to higher silt-clay content in the sediment (Fettweis and Van den Eynde, 2003) and sand extraction typically occurs in regions with medium sands.

Productivity- diversity hypothesis

The relation between productivity and diversity is not unequivocal (Mittelbach *et al.*, 2001). Mostly, a unimodal (i.e. species richness is highest at intermediate levels of productivity), or a positive relationship (i.e. species richness increases with increasing productivity) (Gross and Cardinale, 2007) is found. However, negative relationships have been reported as well (Yount, 1956). The mechanisms that underlie these relationships can be very complex. The positive relation may be explained by mechanisms such as increased survival of rare species or increased abundance of rare resources, while with increasing productivity the diversity may decrease and mechanisms such as a decrease in spatial heterogeneity in the resources and increased competition may take over (Abrams, 1995). Besides, the scales at which those mechanisms operate are also important (Chase and Leibold, 2002). Moreover, the causal relationship between productivity and biodiversity is under discussion: the historical view presumes that productivity drives diversity, however recent evidence shows that diversity can drive production (Cardinale *et al.*, 2009) or mutual effects can be present (Schmid, 2002). It is clear that cause-effect relationships are not distinguishable with our data.

In marine environments, diversity can change according to the sediment: in permeable sediments diversity increased after the deposition of a phytoplankton bloom (Vanaverbeke *et al.*, 2004b). On the other hand, diversity did not change in fine-grained sediments after the sedimentation of phytodetritus (Steyaert, 2003) or decreased with increasing organic input (Armenteros *et al.*, 2010). This negative relationship has been related to the reduced conditions in the sediment resulting in hypoxia, hydrogen sulphide and ammonia, which may have strong negative effects on the nematode community (Gray *et al.*, 2002; Armenteros *et al.*, 2010). In the deep-sea a positive effect of production on the nematode diversity has been observed (Tietjen, 1984; Lambshead *et al.*, 2000; Ingels *et al.*, 2011).

Our data indicates a negative relation between diversity and chl *a* and TSM. This negative relation is observed for all diversity aspects: taxonomic distinctness (Chapter 3), evenness and species richness (Chapter 3 and 4). No interactive effect with the silt-clay or sand fraction could be derived from the data.

Disturbance and productivity

The dynamic equilibrium hypothesis relates disturbance and productivity with species richness: when productivity is low, a negative correlation is found between disturbance and diversity while at a high productivity this correlation is positive (Huston, 1979; Kondoh, 2001). A unimodal disturbance-diversity model is observed at moderate productivity rates. The peak in species richness is a combined effect of a reduction of the competitive species due to the disturbance, but also due to increased number of species able to occupy the niche (niche packing) (Kondoh, 2001). For nematode communities, there is only one study investigating the combined effect of disturbance and productivity: Austen and Widdicombe (2006) found that diversity was highest at low levels of both disturbance and organic enrichment. Moreover, lowest diversity was found at high levels of organic enrichment and no physical disturbance which supports the dynamic equilibrium hypothesis. In the deep-sea, the explanation of depth-diversity pattern has been associated with a non-equilibrium interaction between productivity and disturbance (*in* Lambshead and Boucher, 2003; Ingels *et al.*, 2011). Large-scale physical disturbances however cause a lower local diversity (Lambshead *et al.*, 2001).

Our results from Chapter 2 suggest a positive effect of hydrodynamic properties and a negative effect of organic enrichment. However, the interactive effect of both parameters was not studied. In Chapter 3 the interactions between the nine environmental variables was studied, but current properties or other disturbance related variables were no part of this analysis. Thus, here we are not able to draw conclusions concerning the intermediate dynamic equilibrium hypothesis.

Patchiness, disturbance and biodiversity interrelatedness

Interestingly, patchiness, disturbance and biodiversity may be interrelated: disturbances, such as currents may create patches at large spatial scales, and on a smaller scale biotic interactions may create small scale patches. These processes may produce heterogeneity at various spatial and temporal scales (Richerson *et al.*, 1970). Not only horizontally, but even vertically a mosaic of communities with different diversities and species can be created (Austen *et al.*, 1998; Braeckman *et al.*, 2011). These patches and between patch dynamics may also relate to the intermediate disturbance hypothesis (Wilson, 1990; Collins and Glenn, 1997; Guilini *et al.*, 2011) and diversity at the larger scale may be maximised at some intermediate frequency of patch formation (Abugov, 1982). Grassle & Morse-Porteus (1987) suggested that the deep sea could support large local species richness through the patchy distribution of ephemeral resources in the absence of continuous wide-scale disturbance.

However, in the framework of this research no conclusions can be drawn regarding possible interrelated effects.

Aggregations and spatial autocorrelation

Species distributions are often aggregated due to inherent internal factors (e.g. dispersal, gregarious behaviour, reproduction) as well as due to induced external environmental factors (van Teeffelen and Ovaskainen, 2007). This results in positive spatial autocorrelation where nearby observations are more alike than observations further away. Our analysis, combined with the results of previous research, point in the direction of spatial autocorrelation at both large and small scale.

For the nematode communities, large scale ranges of 42 km for the diversity indices, ES(25) and species richness have been found (Chapter 4). These large ranges can be mainly attributed to the environmental variables (Fig. 4.3). These variables explain about 80% of the variation in the biodiversity (ES(25), Table 4.3) of the nematode community. About 35% of the remaining 20% of the variation is small-scale variation which may be attributed to local variation and patchiness between the replicate samples. Although the environmental variables have a resolution of 200 m, they explain most of the diversity differences between nematode communities. Not all diversity indices are equally strong influenced by the abiotic conditions, the strongest spatial autocorrelation is especially observed in diversity indices representing species richness and evenness. The taxonomic diversity indices show less spatial autocorrelation and the relation with the environmental factors is less pronounced

(Chapter 3) indicating that different types of species communities may be present on a small local scale. Within these large scale patterns, small scale patches with a surface ranging from some square millimetre to some square decimetre may also be discerned (Heip and Engels, 1977; Findlay, 1981; Blanchard, 1990). The factors leading to these small scale patchy distributions are more difficult to establish. On this small local scale, biotic interactions (Reise, 1981; Braeckman *et al.*, 2011), but also small scale differences in the environmental variables (Hogue and Miller, 1981; Sun *et al.*, 1993; Blome *et al.*, 1999) may contribute to these patches.

Metacommunities

Metacommunities are a set of local communities that are linked by dispersal of multiple interacting species on a regional scale (Hubbell, 2001). Depending on the relative importance of environmental heterogeneity (niche concept) and dispersal processes, four types of metacommunities are discerned (Leibold *et al.*, 2004): the species sorting, source-sink dynamics, the neutral model and patch dynamics type. The concept of the metacommunity is mostly theoretical and actual research on metacommunities is impaired since little is known about the individual dispersal capacities of species. Nevertheless, we touch upon two aspects: patch dynamics and species sorting.

Patch dynamics describe species composition between multiple, identical patches, and emphasizes colonisation-competitive ability trade-offs. Here, the species composition in a local community in a sample core is compared with the species composition of sample core originating from the same sampling event. The local species pool forms aggregated communities and no competitive interactions could be discerned in the data. In addition, little is known about the dispersion and colonisation abilities of the species.

Species sorting describes variation in abundance and composition within the metacommunity due to individual species responses to environmental drivers, rather than to competitive interactions. This is based on the niche concept of Hutchinson (1957). In fact, the niche concept is the basis of species distribution modelling: it estimates the environmental niche of the individual species. However, the extrapolation of these estimates to the composition of the metacommunity is impaired due to the limited number of species which could be modelled and the limited models concerning relative abundances of species.

Source-sink models and the neutral model were not treated since data concerning the dispersal capacities and birth and death rates of nematodes is missing.

GENERAL CONCLUSIONS

Modelling

Data assembled from different datasets need careful considerations: in general, sampling campaigns should be developed in such a way that sampling has occurred randomly and in a standardised way. Databases composed from different sources often violate these assumptions and extra care should be taken when analyzing these data: community parameters used to analyze the data should be independent of sampling effort or sampling design. Moreover, **spatial autocorrelation** and **preferential sampling** may be present in the data. These issues are rarely addressed during the same analyses. However, our analyses point out that both aspects **are important**, since they inflate the test statistics and result in falsely accepting a model, while it is in fact not significant (Chapter 5 and 7).

In this thesis different techniques were applied to address these issues. To address spatial autocorrelation we applied four techniques: checking the spatial autocorrelation in the model residuals, applying spatially explicit models, applying geostatistics, and spatially separating test and training sets during cross-validation. Applying spatially explicit models is the most straightforward way to handle this issue. If mapping is desired, the residual spatial autocorrelation may then further be used in the final map by applying geostatistics. However, the latter technique requires a lot of data. When applying other modelling techniques not incorporating spatial techniques, the residual spatial autocorrelation can be tested by calculating Moran's I; or the influence of spatial autocorrelation on the model can be reduced by spatially separating test and training set in cross-validation.

Preferential sampling can be addressed in two ways: (1) preferential sampling of the dependent variable is checked by evaluating the declustered mean or (2) preferential sampling of the environmental data can be discovered by applying null models comparing random models resulting from sampling stations in the total area with random models resulting from stations retrieved from the sampling database. Checking for preferential sampling is essential to identify those models which are significantly different from random.

Another useful way to check the models, is comparing the model outcome with existing knowledge from previous research: complex models may select environmental variables which may explain a part of the variation in the data, but are ecologically irrelevant. In general our models were in accordance with the general knowledge of the taxa under study.

Biodiversity

The null models based on the replicate samples did not reveal negative species interactions. However, the analyses did point out that species tend to aggregate and these aggregations are markedly different between replicate samples, indicating that the nematode communities show a patchy distribution. The factors contributing to this patchiness cannot be derived from the data at hand.

Disentangling the major factors contributing to the current biodiversity patterns of the nematode communities in the Southern Bight of the North Sea is not easy. In the past, competition may have led to the co-evolution of species ('ghost of competition past'). However, this hypothesis cannot be supported by the data at hand. Other hypotheses and relations have been supported with our data: the enigma about the high diversity of the nematode community has been attributed to small-scale heterogeneity (Nielsen *et al.*, 2010) which allows different species to occupy different niches. Based on the data at hand, we indeed find a positive relation between species richness and evenness and the sand fraction and the lowest diversity is correlated with muddy environments. This can be related to the habitat heterogeneity hypothesis on one hand, but also to the oxygen stress in muddy environments on the other hand (Vanaverbeke *et al.*, 2011). The environmental variables related to disturbance seem to be positively correlated with diversity, although the influence is less pronounced compared to the sediment characteristics and these variables may be a proxy for higher sand fractions. Our data indicated a negative relation between productivity and species diversity which may be related to the anoxia resulting from the increase in organic load (Steyaert *et al.*, 1999). The positive relation between productivity and species diversity in permeable sediments (Vanaverbeke *et al.*, 2004b) could not be derived from the data.

Interestingly, patchiness has also been related to an increase in biodiversity: patchiness in the environment and biotic interactions may lead to patchy patterns in the nematode community. Local disturbances may enhance patchiness and thus help increasing local diversity. In this way disturbance has a positive effect through the creation of heterogeneous environments, which in turn allows more species to coexist in a limited area. However, the interrelatedness of these aspects should be further tested in experimental setups.

FUTURE OUTLOOK

- Besides α-diversity, also β- and γ-diversity are important measures to reveal the biodiversity of the marine environment. It is surely a challenge to map the β-diversity based on data from heterogeneous sources. However, maps revealing both the α- and the β-diversity of a region could be important instruments in conservation management.
- Extending the analyses to other taxa could further complete our current knowledge about the diversity and the distribution of these taxa across the Southern Bight of the North Sea. Combining biodiversity maps and HSMs of different taxa can help in establishing vulnerable and valuable regions for conservation.
- The enigma of the high diversity of nematode species remains unresolved. Competition, although not confirmed in this research, may have attributed to the present-day diversity of the nematode community. However, revealing competition as an important factor in diversification is a challenge, and is only possible through carefully monitored experimental setups and evolutionary studies.
- The factors contributing to the local species diversity could be revealed by experiments. These experiments could include factors such as sediment characteristics, oxygen concentration in the sediment at different depths, disturbances and patchiness. Patchiness is a common feature of meiobenthic communities and it could be an important factor in maintaining high biodiversity. Therefore, it could be interesting to investigate on an experimental scale how

patches are formed and how they attribute to the local diversity of the meiobenthic community.

- Introducing environmental variables which directly influence nematode communities, such as oxygen penetration depth, may further improve the models. Moreover, modelling seasonal fluctuations in the benthic communities based on seasonal environmental data may further enhance our current understanding of the benthic ecosystem.
- Geostatistics requires a high number of observations to allow reliable mapping. The diversity maps (Chapter 4) show patches in the Northern part of the region and a higher number of sampling data in the northern region could help in improving the maps. The distance between the samples would be preferably smaller than the range of the variograms (Fig. 4.3). More specifically, this would imply a sampling distance smaller than 10 km for the estimation of ES(25).
- Maxent has the ability to project the species' models to future environmental scenarios. Thus, the knowledge of future concentrations of chlorophyll *a* and total suspended matter, could be used to predict how species compositions could change under future scenarios. Especially, data covering the changes induced by climate change could reveal the potential impact of climate change on the nematode community. However, this can only be done if the model does not extrapolate beyond the range of the environmental data used to build the model.
- Investigating the biotic effect of macrobenthic species (i.e. habitat engineering species) on the diversity and presence of nematode communities and the knowledge of the distribution of these macrobenthic species may further help in improving our current knowledge.